1997

Semiparametric Maximum Likelihood Estimation of Nonlinear Regression Models and Monte Carlo Evidence

Jian Yang

ISSN:0318-725X
ISBN:0-7714-2033-1

RESEARCH REPORT 9713

Semiparametric Maximum Likelihood Estimation of Nonlinear Regression Models and Monte Carlo Evidence

by

Jian Yang

August 1997

Department of Economics
Social Science Centre
University of Western Ontario
London, Ontario, Canada
N6A 5C2
econref@sscl.uwo.ca

Semiparametric Maximum Likelihood Estimation of Nonlinear Regression Models and Monte Carlo Evidence

Jian Yang *
Department of Economics
University of Western Ontario
London, Ontario, N6A 5C2
yang@sscl.uwo.ca

August 1997

ABSTRACT

This paper studies adaptive estimation of nonlinear regression models with i.i.d. error terms. Previously in the literature, the adaptive maximum likelihood estimator (AMLE) has been proposed only when the densities of the error terms are assumed to be symmetric. AMLE is a one-step estimator and utilizes outer-product of gradient (OPG) matrix to estimate the information matrix, which could cause serious finite sample problems. I propose semiparametric maximum likelihood estimator (SMLE) for the nonlinear regression models with or without the symmetry condition. SMLE is defined by maximizing the nonparametric log-likelihood function constructed using the residuals from an initial root-n consistent estimate. I show that SML estimators are adaptive (asymptotically efficient) for all the adaptively estimable parameters in nonlinear regression models assuming or without the symmetry condition. Further, SML estimators are generally consistent for all the structural parameters even without the symmetry assumption of the error distribution.

Monte Carlo studies show that, in the symmetric cases, SMLE performs the best and OLSE the worst for all the parameters based on the root MSE and interquartile range standards. The more the distance of the true distribution differs from the normal distribution, the greater the relative efficiency gains SMLE and AMLE achieve over OLSE. In the asymmetric cases, SMLE performs the best for the adaptively estimable parameters. The more the distance of the true distribution differs from the normal distribution, the greater the relative efficiency gain SMLE achieves over OLSE.

We also find that the t-ratio test computed from the AML procedure tends to over-reject the null hypothesis. In contrast, the t-ratio tests computed from the SML procedure and the OLS procedure work quite well. The t-ratio test computed from the SML procedure has the greatest power, followed by the t-ratio test computed from the AML procedure in the symmetric cases and the t-ratio test computed from the OLS procedure in the asymmetric case.

* This paper is drawn from the first two chapters of my thesis at Queen's University. I would like thank my supervisors, Russell Davidson and James MacKinnon, for their patience and guidence. I also acknowledge helpful comments from Gregor Smith and all the participants, especially Joel Horowitz and Jeff Racine, in the Canadian Econometric Study Group meeting, 1997. All remaining errors are mine. I gratefully acknowledge financial support from Queen's University and the Social Sciences and Humanities Research Council of Canada.

1. Introduction

One of the most common estimation techniques is **maximum likelihood** (**ML**) estimation. It has been widely used in many fields. Part of the attraction of ML theory is that it offers estimators and test procedures that are technically almost universally applicable, provided one has a reasonably precise model. More importantly, ML theory offers estimators which are consistent, asymptotically normal and asymptotically efficient under appropriate regularity conditions.

A fundamental assumption underlying these properties of ML estimators is that the stochastic law which determines the behavior of the phenomena investigated is known to lie within a specified parametric family of probability distributions (the model). In other words, the data generating process, or **DGP**, is assumed to be fully specified by a parametric probability model.

In theoretical as well as in empirical research, the assumption that the error terms are normally distributed has been widely used to implement ML estimation. ML estimators constructed under the normality assumption are called Gaussian ML estimators in the literature. However, considerable evidence has been provided to reject the assumption of normality, for example, in research on financial time series. The unconditional distribution of returns to financial assets exhibits much fatter tails than a normal distribution. Moreover, skewness and leptokurtosis are found in the conditional distribution as well (see Bollerslev: 1987 and Peruga: 1988).

Regarding the empirical evidence of a violation of the normality assumptions, it is natural to ask what happens to the properties of the MLE when the model is mis-specified by assuming normality of the error terms. To answer this question, Gaussian quasi-maximum-likelihood estimation (**QMLE**) was studied by White (1982). He found that Gaussian QMLE inherits, under regularity conditions, the consistency and asymptotic normality properties of MLE. However, Gaussian QMLE is no longer asymptotically efficient. It has a loss of efficiency due to falsely assuming normality.

There are two alternative approaches used in the literature to improve efficiency and obtain more precise estimates. The first approach shares the spirit of Gaussian MLE as a parametric approach. It assumes that the error terms follow a family of parametric distributions other than the normal distribution specifically to fit different data sets.

In principle, a researcher can always search through a large number of non-normal parametric distributions until a good fit is found for a particular data set, but the computation would be intensive. Moreover, when the likelihood is assumed to be a specific non-Gaussian distribution, the so-constructed non-Gaussian QMLE is generally not consistent if the true error distribution is not the assumed non-Gaussian distribution (see Newey and Steigerwald: 1994). Lastly, an estimator constructed from a distribution found in this way suffers from pre-test bias.

An alternative approach is the so-called semiparametric estimation method. Here, the models are specified quite differently from parametric models. Although we still specify the conditional first and second moments parametrically, we do not require parametric specifications of the error distributions. Instead of specifying an unfounded parametric distribution function, we only assume that the distribution function of the error terms is a member of a general family. More technically, we assume that the distribution is nonpara-

metrically specified by an infinite number of parameters rather than by a finite number of parameters in a parametric model. We call these types of models "semiparametric models". Since the nuisance parameters, which characterize the conditional distributions, are infinite in number, we are unable to obtain enough information for the construction of consistent estimators of them. But we are able, as we will find, to obtain enough information for constructing consistent estimators of structural parameters, and the so-constructed estimators are called semiparametric estimators.

Several issues arise associated with the semiparametric approach. The first is whether, and under what conditions, we are able to construct asymptotically efficient estimators by the semiparametric approach. This is labelled "adaptive estimation" in the literature. The second issue that needs to be addressed is how to construct asymptotically efficient estimators of structural parameters when the adaptive conditions are satisfied. The third issue concerns the finite-sample performance and practical usefulness of these adaptive estimators.

In this paper, we consider nonlinear regression models with i.i.d. error terms and also the special case of these models in which there is a free intercept. Manski (1984) showed that all the structural parameters in general nonlinear regression models were adaptively estimable when the error distributions were assumed to be symmetric. In nonlinear regression models with a free intercept, he found that all the structural parameters except the intercept were adaptively estimable even without the symmetry assumption. Manski (1984) proposed the so-called Adaptive Maximum Likelihood Estimators (AMLE) for nonlinear regression models under the symmetry assumption, where the outer-product of the gradient (OPG) estimate of the information matrix was utilized. However, he failed to provide any type of adaptive estimators for the adaptively estimable parameters in nonlinear regression models with a free intercept without the symmetry condition. We argue that, although AMLE is adaptive for nonlinear regression models under the symmetry condition, it could behave poorly in small samples. This is simply because it is a one-step estimator and utilizes the OPG matrix to estimate the information matrix, and the estimator of the information matrix constructed using the OPG matrix behaves poorly in small samples. We suggest an alternative approach to construct adaptive estimators, where the estimators are obtained by maximizing the nonparametric log-likelihood function constructed by using the residuals from the initial $\sqrt{n}$-consistent estimates.[1] The estimators in this way are called semiparametric maximum likelihood estimators, or **SMLE**. Using the SML approach, we are able to construct adaptive estimators not only for all the structural parameters in nonlinear regression models under the symmetry assumption but also for the adaptively estimable parameters in nonlinear regression models with a free intercept without the symmetry condition. Additionally, we prove that SML estimators are generally consistent for all structural parameters in any type of nonlinear regression models with i.i.d. error terms, with or without the symmetry condition. Therefore we could apply SML estimation method to empirical studies more confidently even when the symmetry

[1] Engle and González-Rivera (1991) have adopted a similar approach to estimate ARCH models. They focus on the finite sample performances of the estimators constructed in this way rather than on the construction of adaptive estimators. Besides, there are no trimming parameters used to construct their estimators.

condition is not clearly justified. Finally, as we will find from the Monte Carlo results, SMLE promises much better small sample performance than AMLE and the ordinary least squares estimator in terms of relative efficiency and inference performance.

The rest of the paper is organized as follows. In section 2, we conduct a literature review of adaptive estimation and related topics. In section 3, we present our model and define SMLE of the model. Section 4 examines the asymptotic properties of SMLE for the general nonlinear regression model under the symmetry condition. The asymptotic properties of SMLE for nonlinear regression models with a free intercept are discussed in Section 5. Section 6 conducts Monte Carlo studies.

2. Literature Review

In this section, we perform a literature review on adaptive estimation and related topics. We first discuss the issue of adaptive estimation, and describe the necessary conditions for adaptation. We then discuss the issue of "contiguity", and at final, briefly review the approach used in the literature to construct adaptive estimators.

2.1. The Necessary Conditions for Adaptation

An estimator may be termed adaptive if its computation incorporates a data-based procedure for learning unknown features of the error distribution and if this estimator is asymptotically as efficient as the best that would be attainable were the distribution known (see Manski, 1984).

The problem of adaptive estimability was originally posed by Stein (1956) in the following framework. Assume that a density function g_0 is known to be from a family of distributions characterized by a finite parameter vector $(\beta_0, \eta_0) \in \mathcal{B} \otimes \mathcal{T}$, where $\mathcal{B} \subset \boldsymbol{R}^k$ and $\mathcal{T} \subset \boldsymbol{R}^j$, β_0 are the parameters of interest, and η_0 are the nuisance parameters. Let $I^{-1}(\beta_0)$ be the inverse of the Fisher's information matrix associated with estimation of β_0 given knowledge of η_0, which is the best precision given η_0. Let $I(\beta_0, \eta_0)$ be the information matrix associated with joint estimation of (β_0, η_0). If η_0 is unknown, the best attainable precision of the estimation of β_0 is the upper left $k \times k$ sub-matrix of $I^{-1}(\beta_0, \eta_0)$, which exceeds $I^{-1}(\beta_0)$ by a non-negative semi-definite matrix, which is null if $I(\beta_0, \eta_0)$ is block diagonal. This block-diagonality is the condition that knowledge of η_0 is asymptotically irrelevant to the estimation of β_0.

More generally, the distribution is known from a specific function space Φ with infinite dimension. Stein then concluded that given the prior restriction of g_0 to Φ, β_0 cannot be estimated adaptively if there exists any finite parametric family such that $g_0 \in (g_{\eta_0}, \eta_0 \in \boldsymbol{R}^j) \subset \Phi$ and the upper right $k \times j$ sub-matrix of $I(\beta_0, \eta_0)$ is non-null, in other words, $I(\beta_0, \eta_0)$ is not block diagonal.

A less stringent situation is to consider the adaptation of a given sub-vector of β_0 instead of the whole vector. Let $\beta_0 = [\beta_{10}^\top, \beta_{20}^\top]^\top$, $\beta_{10} \in \boldsymbol{R}^{k_1}$, $\beta_{20} \in \boldsymbol{R}^{k_2}$, and $k_1 + k_2 = k$. The information matrix associated with joint estimation of $I(\beta_0, \eta_0)$ is

$$I(\beta_0, \eta_0) = \begin{pmatrix} I(\beta_1, \beta_2) & I_{\beta\eta} \\ I_{\eta\beta} & I_{\eta\eta} \end{pmatrix} = \begin{pmatrix} I_{\beta_1\beta_1} & I_{\beta_1\beta_2} & I_{\beta_1\eta} \\ I_{\beta_2\beta_1} & I_{\beta_2\beta_2} & I_{\beta_2\eta} \\ I_{\eta\beta_1} & I_{\eta\beta_2} & I_{\eta\eta} \end{pmatrix}.$$

The smallest possible asymptotic variance for an estimator of β_{10} with known η_0 is the $k_1 \times k_1$ upper left sub-matrix of $I^{-1}(\beta_1, \beta_2)$; and the smallest possible asymptotic variance for an estimator of β_{10} with unknown η_0 is the $k_1 \times k_1$ upper left sub-matrix of $I^{-1}(\beta_0, \eta_0)$. Stein (1956) proved that they are equal to each other if and only if

$$I_{\beta_1\eta} - I_{\beta_1\beta_2} I^{-1}_{\beta_2\beta_2} I_{\beta_2\eta} = 0. \tag{2.1}$$

The necessary conditions for adaptation have been examined for a number of models. These include linear regression models with i.i.d. error terms (Bickel: 1982), nonlinear regression models with i.i.d. error terms (Manski: 1984), stationary ARMA process (Kreiss: 1987) and linear ARCH models (Linton: 1993). The adaptive conditions are found to be satisfied for some or all of the structural parameters in different types of models under or without the symmetry assumption of the error distributions.

Manski (1984) considered a non-linear regression model of the following form

$$y_i = h(\boldsymbol{x}_i, \beta_0) + u_i, \tag{2.2}$$

where $y_i \in \boldsymbol{Y} \subset \boldsymbol{R}^1$, $\boldsymbol{x}_i \in \boldsymbol{X} \subset \boldsymbol{R}^m$ are observable, $u_i \in \boldsymbol{R}^1$ is unobservable. The samples $\boldsymbol{x}_i$, $i = 1, ..., n$ are *i.i.d.* with common distribution F and density function f, $u_i \sim$ *i.i.d.* g_0, where f and g_0 are unknown. When g_0 was symmetric with respect to its argument, Manski (1984) found that β_0 in (2.2) was adaptively estimable. The intuition is as follows. When g_0 is symmetric with respect to its argument, the score function of β_0 is anti-symmetric with respect to the residuals, since g_0' is anti-symmetric, where g_0' is as usual the first order derivative of g_0. However, for any nuisance parameter η_0, which characterizes g_0, the score function of η_0 is symmetric, since $\partial g_0 / \partial \eta_0$ is symmetric if g_0 is symmetric for all η_0. Then the two score functions are orthogonal to each other, which simply means that the information matrix $I(\beta_0, \eta_0)$ is block diagonal.

The symmetry condition is required for the adaptation of all structural parameters in a general nonlinear regression model (2.2). However, this fairly strict assumption is not necessary for all the sub-families of nonlinear models. Manski (1984) studied a non-linear regression model with free intercept

$$y_i = \alpha_0 + h_1(\boldsymbol{x}_i, \theta_0) + u_i. \tag{2.3}$$

He found that θ_0 in this model satisfied condition (2.1) as long as u_i was i.i.d.. θ_0 is then adaptively estimable, although α_0 is not in this case.

2.2. The Construction of Adaptive Estimator

The first construction of adaptive estimation was done almost fifteen years after Stein's paper for the simplest regression problem, namely the location parameter problem in which

$$y_i - \beta_0 = u_i, \tag{2.4}$$

where y_i, β_0, and u_i are scalars, u_i has density g_0, and a random sample of values of y are observed. The most general results were achieved by Beran (1974), who showed that if g_0 was known only to be symmetric around zero, one could construct an estimator for β_0 whose asymptotic variance was equal to the best that would be attainable were g_0 known. Thus, given symmetry of g_0, β_0 was adaptively estimable.

Because Beran's approach involves the construction of adaptive rank estimates, it does not lend itself to applications to more complex estimation problems. However, Stone (1975) reported an alternative constructive proof that a location parameter could be estimated adaptively, given symmetry of g_0.

Stone's construction has the following steps:

(1) Compute $\tilde{\beta}_n$, any estimate for β_0 which does not use knowledge of g_0 and which is $\sqrt{n}$-consistent whenever g_0 is a symmetric density.

(2) Calculate the residuals $\tilde{u}_{in} = y_i - \tilde{\beta}_n$, $i = 1, ..., n$.

(3) Use the residuals to form a nonparametric estimate $\tilde{g}_n$ of the density g_0. Stone chose a particular trimmed kernel estimate.

(4) Acting as if the density estimate is the true density, take one linearized Newton-Raphson type step from $\tilde{\beta}_n$,

$$\hat{\beta}_n = \tilde{\beta}_n + [I^n(\tilde{\beta}_n)]^{-1} S^n(\tilde{\beta}_n), \tag{2.5}$$

where

$$I^n(\tilde{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{g}_n'}{\tilde{g}_n}\right)^2 (y_i - \tilde{\beta}_n),$$

and

$$S^n(\tilde{\beta}_n) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{g}_n'}{\tilde{g}_n}\right)(y_i - \tilde{\beta}_n),$$

where $y_i - \tilde{\beta}_n$ in these two equations is an argument. They are respectively nonparametric estimations of the information matrix and the score vector of β_0 based on $\tilde{\beta}_n$, and $I^n(\tilde{\beta}_n)$ is estimated by the sample mean of the outer product of the gradient based on $\tilde{\beta}_n$. In the idealized situation with known density g_0, we denote the information matrix of β_0 as $I(\beta_0, g_0)$. Then the estimator $\hat{\beta}_n$ is adaptive for β_0 if and only if

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, I^{-1}(\beta_0, g_0)). \tag{2.6}$$

Usually, the kernel method with Gaussian kernel has been used to form a nonparametric density estimator in the adaptive estimation literature. Define the convolution of g_0 and ϕ_{λ_n} as

$$g_{\lambda_n}(u) = g * \phi_{\lambda_n}(u) = \int_{-\infty}^{+\infty} \phi_{\lambda_n}(u - z) g(z) dz. \tag{2.7}$$

We consider the idealized situation in which we are able to access the true disturbance terms. The density estimator is then of the following form.

$$g_{e\lambda_n}(u) \equiv n^{-1} \sum_{i=1}^{n} \phi_\lambda(u - u_i). \tag{2.8}$$

To construct a type of AML estimator, one needs to estimate the score function and the information matrix for the parameters. For example, in the simplest regression model,

$$y_i = \beta_0 + u_i, \ i = 1, \cdots, n,$$

we have one location parameter β_0. The density function of u_i is g_0, then the loglikelihood contributed by observation i is simply

$$\ell(y_i; \beta_0) = \log(g_0(y_i - \beta_0)).$$

The score function of β_0 for observation i is

$$S_i(\beta_0, g_0) = \frac{\partial \ell}{\partial \beta}(y_i; \beta_0) = -\frac{g_0'}{g_0}(y_i - \beta_0).$$

The OPG estimate of the information matrix is

$$\frac{1}{n} \sum_{i=1}^{n} \left[\frac{g_0'}{g_0}(y_i - \beta_0) \right]^2.$$

Appropriate estimation of the information matrix and the score function requires suitable consistent estimation of the score function g_0'/g_0 rather than that of the density function.

Stone (1975) proposed a family of trimmed score function estimators of g_0'/g_0. However, his approach is rather burdensome to compute: it is therefore not discussed here. Bickel (1982) was able successfully to provide different trimming rules, which were widely used to construct AMLE. In the following, we briefly describe why it is necessary to use trimming and how the trimming parameters work in the construction of AMLE.

Let $S_{n\lambda_n}$ be an estimator for the score function g_0'/g_0. Then $S_{n\lambda_n}$ is required to satisfy the mean square convergence condition

$$\int \left[S_{n\lambda_n}(u) - \frac{g_0'}{g_0}(u) \right]^2 g_0(u) du \to o_p(1), \tag{2.9}$$

as $n \to \infty$. A straightforward nonparametric estimate of the score function g_0'/g_0 is

$$\frac{g_{e\lambda_n}'}{g_{e\lambda_n}}(u).$$

However, this estimator does not naturally satisfy the mean square convergence condition of (2.9). This is because although $g_{e\lambda_n}(u)$ and $g_{e\lambda_n}'(u)$ have well defined properties.

$g'_{e\lambda_n}/g_{e\lambda_n}(u)$ is not well behaved everywhere. When $g_{e\lambda_n}$ or $g'_{e\lambda_n}$ is very small, we will have

$$\frac{g'_{e\lambda_n}}{g_{e\lambda_n}}(u) = \frac{1}{0} \text{ or } \frac{0}{0}.$$

Bickel (1982) then suggested using a trimmed score function defined as

$$S_{n\lambda_n}(u) = \begin{cases} \frac{g'_{e\lambda_n}}{g_{e\lambda_n}}(u) & \text{if } |u| \leq e_n, g_{e\lambda_n}(u) \geq d_n, \text{ and } \left|\frac{g'_{e\lambda_n}}{g_{e\lambda_n}}(u)\right| \leq c_n, \\ 0 & \text{otherwise,} \end{cases} \tag{2.10}$$

where e_n, d_n and c_n are positive parameters chosen to control trimming. The value of the nonparametric density function $g_{e\lambda_n}(u)$ is controlled by d_n such that it is bounded away from 0. The value of the nonparametric score function of $g_{e\lambda_n}(u)$, $g'_{e\lambda_n}/g_{e\lambda_n}(u)$, is controlled by c_n to be bounded away from ∞. And u is bounded by e_n, in the domain of integration in (2.9). It is clear that when $n \to \infty$, we need $e_n \to \infty$, $d_n \to 0$, and $c_n \to \infty$ to have well defined asymptotic properties. Bickel (1982) provided some specific trimming rules to control the convergence of $S_{n\lambda_n}(u)$ to $g'_0/g_0(u)$. These trimming rules are $n^{-1}e_n\lambda_n^{-3} \to 0$ and $\lambda_n c_n \to 0$ as $n \to \infty$.

2.3 Contiguity

In the above discussion, we only consider the idealized situation in which true error terms are observable. Unfortunately, we are not able to access true error terms. What we have in order to construct nonparametric density functions are residuals from initial root-n consistent estimates. The natural question would be whether these residuals can replace disturbances to construct nonparametric density functions. The following argument relies on an idea of LeCam (1960) and on a theorem of Hajek and Sidak (1967). LeCam (1960) introduced the concept of "contiguity" of two sequences of probability measures. Consider two sequences of absolutely continuous probability measures on $\boldsymbol{R}^n$ $\{P_n, Q_n\}$ with $\{p_n, q_n\}$ as the associated density sequences.

Definition 2.1.

If for any sequence of events $A_n \subset \boldsymbol{R}^n$, $n = 1, ..., +\infty$,

$$[P_n(A_n) \to 0] \Rightarrow [Q_n(A_n) \to 0] \tag{2.11}$$

holds, then the densities q_n are said to be contiguous to the densities p_n, where $dP_n = p_n d(\mu_n)$, $dQ_n = q_n d(\mu_n)$, and $d(\mu_n)$ is the Lebesgue Measure in $\boldsymbol{R}^n$.

Hajek and Sidak (1967) provide a useful result on contiguity. Assume that z_i, $i = 1, ..., +\infty$ is a sequence of independent scalar random variables and that the joint density of z_i, $i = 1, ..., n$ is

$$p_n = \prod_{i=1}^{n} g(z_i - \mu_{in}), \tag{2.12}$$

where $0 < I(g) < \infty$,

$$I(g) \equiv \int \left(\frac{g'}{g}\right)^2 (z) g(z) dz,$$

and μ_{in}, $i = 1, ..., n$ are a set of location parameters. Let $\bar{\mu}_n = n^{-1} \sum \mu_{in}$ and define

$$q_n = \prod_{i=1}^{n} g(z_i - \bar{\mu}_n). \tag{2.13}$$

Lemma 2.1.

Assume that as $n \to +\infty$,

$$\max_{i=1,...,n} (\mu_{in} - \bar{\mu}_n)^2 \to 0, \tag{2.14}$$

and

$$\sum_{i=1}^{n} (\mu_{in} - \bar{\mu}_n)^2 \to \rho, \tag{2.15}$$

where $0 < \rho < \infty$. *Then* $\{q_n\}$ *is contiguous to* $\{p_n\}$.

Manski (1984) generalizes this result to the non-linear regression model (2.2), where the samples $\boldsymbol{x}_i$, $i = 1, ..., n$ are *i.i.d.* with common distribution F and density function f, $u_i \sim$ *i.i.d.* g_0. He also made the following assumption.

Assumption 2.1

$E \sup_{\boldsymbol{\beta} \in \mathcal{B}} |\ell(y; \boldsymbol{x}, \boldsymbol{\beta})| < +\infty$, $\ell(y; \boldsymbol{x}, \boldsymbol{\beta})$ *is a twice continuously differentiable function of* $\boldsymbol{\beta}$, *where* $\boldsymbol{\beta} \in \mathcal{B} \subset \boldsymbol{R}^k$ *and* $\ell(y; \boldsymbol{x}, \boldsymbol{\beta})$ *is the log-likelihood function of* y *given observation* $\boldsymbol{x}$ *under* $\boldsymbol{\beta}$. $E[(\frac{\partial h(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})(\frac{\partial h(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal})]$ *is finite positive definite matrix, given that* $|\boldsymbol{\beta} - \boldsymbol{\beta}_0| < \varepsilon$ *with some* $\varepsilon > 0$.

Lemma 2.2: (Manski)

For model (2.2) *with Assumption* 2.1, *and given any sequence of* $\sqrt{n}$*-consistent estimates* $\boldsymbol{\beta}_n$, $n = 1, ..., +\infty$, *the sequence of densities of the residuals is contiguous to the sequence of densities of the disturbances, almost surely in* $\boldsymbol{x}$.

With Lemma 2.2, we have the following result (see Bickel, 1982 and Manski, 1984).

Lemma 2.3:

Given the same assumptions as in Lemma 2.2, *if the estimator of the score function* g_0'/g_0, $S_{n\lambda_n}(u)$ *in* (2.10), *satisfies the mean square convergence condition* (2.9), *then* $\tilde{S}_{n\lambda_n}(u)$ *continues to satisfy this condition, where* $\tilde{S}_{n\lambda_n}(u)$ *is computed in the same way as* $S_{n\lambda_n}(u)$ *by replacing the true error terms with the residuals.*

The basic intuition of Lemma 2.3 is as follows. Given any small number ε, define a sequence of events $\{A_n\}$ by

$$A_n = \left[(u_1, \cdots, u_n) \in \boldsymbol{R}^n : \int \left[S_{n\lambda_n}(u) - \frac{g_0'}{g_0}(u)\right]^2 g_0(u) du > \varepsilon\right].$$

If $P(A_n) \to 0$ under the probability measures of the disturbances, then $P(A_n) \to 0$ under the probability measures of the residuals.

By these results, we can simply use $\sqrt{n}$ consistent residuals to replace the true error terms to construct adaptive maximum likelihood estimators and draw the same conclusions.

In the construction of the AMLE, we only need to show the mean square convergence of $\tilde{S}_{n\lambda_n}(u)$ to $g_0'(u)/g_0(u)$. It is justified by Lemma 2.3 that this condition holds as long

as $S_{n\lambda_n}(u)$ converges to $g_0'(u)/g_0(u)$. Actually, we can obtain some other results, which we will find useful in the following discussion.

In model (2.2) under Assumption 2.1 and the assumption that $\int |g_0^{(2)}(u)|du < \infty$, we define $\ell_{n\lambda_n}(u)$ as a nonparametric estimator of the log-density, $\log(g_0(u))$, $H_{n\lambda_n}(u)$ as a nonparametric estimator of the function, $g_0^{(2)}(u)/g_0(u)$, where $\ell_{n\lambda_n}(u)$ and $H_{n\lambda_n}(u)$ are computed based on the true disturbances $u_1, \cdots, u_n$. $\tilde{\ell}_{n\lambda_n}(u)$ and $\tilde{H}_{n\lambda_n}(u)$ are computed respectively in the same way as that of $\ell_{n\lambda_n}(u)$ and $H_{n\lambda_n}(u)$ with the disturbances replaced by the $\sqrt{n}$ consistent residuals. We then have the following corollary of Lemma 2.2.

Corollary 2.1:

Given the same assumptions as in Lemma 2.2, and that $\int |g_0^{(2)}(u)|du < \infty$, *if* $\ell_{n\lambda_n}(u)$ *and* $H_{n\lambda_n}(u)$ *satisfy the mean absolute convergence conditions*

$$\int |\ell_{n\lambda_n}(u) - \log(g_0(u))| g_0(u)du \to o_p(1),$$

and

$$\int \left| H_{n\lambda_n}(u) - \frac{g_0^{(2)}(u)}{g_0(u)} \right| g_0(u)du \to o_p(1),$$

then $\tilde{\ell}_{n\lambda_n}(u)$ *and* $\tilde{H}_{n\lambda_n}(u)$ *continue to satisfy these conditions.*

By these results, we can simply use root-n consistent residuals to replace the true error terms to construct adaptive maximum likelihood estimators and draw the same conclusions.

2.3 AMLE of Nonlinear Regression Models

We have so far only introduced the issue of adaptation, the necessary conditions for adaptation, and some techniques used in the construction of AMLE. However, we have not had a chance to see more closely how AMLE is constructed, and in particular, what kind of role the symmetry condition plays.

For an illustration, we consider a nonlinear regression model described in (2.2). Manski (1984) found that β_0 is adaptively estimable when g_0 is symmetric and he constructed an adaptive estimator of β_0 using the AML approach.

For any given β, the contribution to the log-likelihood by observation i is

$$\ell(y_i; \boldsymbol{x}_i, \beta) \equiv \ell(y_i - h(\boldsymbol{x}_i, \beta)) = \log[g_0(y_i - h(\boldsymbol{x}_i, \beta))]. \tag{2.16}$$

The score function for observation $i = 1, \cdots, n$ under g_0 is

$$S_i(\beta, g_0) \equiv -\frac{\partial h(\boldsymbol{x}_i, \beta)}{\partial \beta} \frac{g_0'}{g_0}(y_i - h(\boldsymbol{x}_i, \beta)), \tag{2.17}$$

and the information matrix of β_0 under g_0 is

$$I(\beta_0, g_0) = E\left[\left(\frac{\partial h(\boldsymbol{x}, \beta_0)}{\partial \beta}\right) \left(\frac{\partial h(\boldsymbol{x}, \beta_0)}{\partial \beta^\top}\right) \right] E\left[\left(\frac{g_0'}{g_0}(u) \right)^2 \right]. \tag{2.18}$$

The sample mean of the score function and the OPG matrix are respectively

$$S^n(\beta, g_0) \equiv \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial h(x_i,\beta)}{\partial \beta}\frac{g_0'}{g_0}(y_i - h(x_i,\beta)), \tag{2.19}$$

and

$$I^n(\beta_0, g_0) = \frac{1}{n}\sum_{i=1}^{n}\left[(\frac{\partial h(x_i,\beta_0)}{\partial \beta})(\frac{\partial h(x_i,\beta_0)}{\partial \beta^\top})\right]\left[\left(\frac{g_0'}{g_0}(y_i - x_i\beta_0)\right)^2\right]. \tag{2.20}$$

Most researchers in the adaptive estimation literature started from the discussion of Local Asymptotic Normality (LAN), where they assumed ideally that g_0 was known. Provided minimal smoothness restrictions, see among others LeCam (1969) and Bickel (1982), it is shown that there exists a sequence of estimators $\{\hat{\beta}_n\}$ of β_0 such that whenever $n^{1/2}|\beta_n - \beta_0| \leq M$ for all n, $M < \infty$,

$$n^{1/2}(\hat{\beta}_n - \beta_n) \xrightarrow{d} N(0, I^{-1}(\beta_0, g_0)). \tag{2.21}$$

Instead of starting from primitive conditions, Manski (1984) simply assumed that the estimation problem is sufficiently regular that

$$\hat{\beta}_n = \beta_n + \{I^n(\beta_n, g_0)\}^{-1} S^n(\beta_n, g_0) \tag{2.22}$$

satisfies condition (2.21). (2.22) is similar to the one step efficient estimator; the difference is that the one step efficient estimator usually uses the sample mean of the Hessian matrix rather than $I^n(\beta_n)$.

Now we return to the situation with unknown g_0. The problem is whether we are still able to find $\hat{\beta}_n$ satisfying (2.21) without the knowledge of g_0. A simple definition of an adaptive estimator can be found in Bickel (1982).

Definition 2.2

A sequence of estimators $\{\hat{\beta}_n\}$ is adaptive if and only if, for every regular (β_0, g_0),

$$n^{1/2}(\hat{\beta}_n - \beta_n) \to N(0, I^{-1}(\beta_0, g_0)) \text{ in distribution,} \tag{2.23}$$

whenever $n^{1/2}|\beta_n - \beta_0|$ stays bounded.

From Definition 2.2, if an estimator $\hat{\beta}_n$ is adaptive, it is then efficient in the usual sense. This is simply because $\{\beta_n\}$ is a nonstochastic sequence and β_n converges to β_0 in the order of $\sqrt{n}$.

In the described nonlinear regression model, the density function g_0 is unknown. Suppose we have a nonstochastic sequence β_n such that $n^{1/2}|\beta_n - \beta_0|$ stays bounded, we denote the residuals $\tilde{u}_{in} = y_i - h(x_i, \beta_n)$, $i = 1, ..., n$. These residuals will be used to construct nonparametric density functions and nonparametric score functions. Since g_0 here is assumed to be symmetric, it is desirable to use a symmetric estimator of the density

function. However, the density estimator defined in (2.8) is not symmetric. Manski (1984) suggested using

$$g_{e\lambda_n} = \frac{1}{2}[g_{e\lambda_n}(u) + g_{e\lambda_n}(-u)]. \tag{2.24}$$

For notational simplicity, we still use $g_{e\lambda_n}$ to denote the symmetric density estimator. In the rest of this study, if we assume the symmetry of g_0, we define $g_{e\lambda_n}$ by (2.24); otherwise, we define it by (2.8).

Using the residuals $\tilde{u}_{in}$, $i = 1, ..., n$ instead of u_i $i = 1, ..., n$, we can compute a pseudo nonparametric density function $\tilde{g}_{e\lambda_n}(y_i - \boldsymbol{x}_i\beta_n)$ according to (2.24) and a pseudo nonparametric trimmed score function $S_i(\beta_n, \tilde{g}_{e\lambda_n})$ for $i = 1, \cdots, n$. We then have the sample mean of the score function estimate and the information matrix estimate as respectively

$$S^n(\beta_n, g_{e\lambda_n}) \equiv \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial h(\boldsymbol{x}_i, \beta_n)}{\partial \beta}\frac{g'_{e\lambda_n}}{g_{e\lambda_n}}(y_i - h(\boldsymbol{x}_i, \beta_n)), \tag{2.25}$$

and

$$I^n(\beta_n, g_{e\lambda_n}) = \frac{1}{n}\sum_{i=1}^{n}\left[(\frac{\partial h(\boldsymbol{x}_i, \beta_n)}{\partial \beta})(\frac{\partial h(\boldsymbol{x}_i, \beta_n)}{\partial \beta^\top})\right]\left[\left(\frac{g'_{e\lambda_n}}{g_{e\lambda_n}}(y_i - \boldsymbol{x}_i\beta_n)\right)^2\right], \tag{2.26}$$

where the three conditions of trimming in (2.10) are satisfied.

With these results at hand, Manski (1984) constructed the estimator in the fashion of (2.5).

$$\hat{\beta}_n = \beta_n + \{I^n(\beta_n, g_{e\lambda_n})\}^{-1} S^n(\beta_n, g_{e\lambda_n}) \tag{2.27}$$

This is called the AMLE. To show that the AMLE is adaptive, one mainly needs to show that

$$\int \left[\sqrt{n}(S^n(\beta_n, g_{e\lambda_n}) - S^n(\beta_n, g_0))\right]\left[\sqrt{n}(S^n(\beta_n, g_{e\lambda_n}) - S^n(\beta_n, g_0))\right]^\top g_0(u)du \to o_p(1) \tag{2.28}$$

The left-hand side of (2.28) can be rewritten as

$$\int \frac{1}{n}\sum_{i,j}\Bigg\{\left[(\frac{\partial h(\boldsymbol{x}_i, \beta_n)}{\partial \beta})(\frac{\partial h(\boldsymbol{x}_j, \beta_n)}{\partial \beta^\top})\right]$$

$$\left[\left(\frac{g'_{e\lambda_n}}{g_{e\lambda_n}} - \frac{g'_0}{g_0}\right)(y_i - \boldsymbol{x}_i\beta_n)\right]\left[\left(\frac{g'_{e\lambda_n}}{g_{e\lambda_n}} - \frac{g'_0}{g_0}\right)(y_j - \boldsymbol{x}_j\beta_n)\right]\Bigg\} g_0(u)du, \tag{2.29}$$

Given that g_0 and $g_{e\lambda_n}$ are symmetric, then g'_0 and $g'_{e\lambda_n}$ are anti-symmetric. The cross-products for $i \neq j$ in (2.29) drop out. It thus becomes

$$\int \frac{\partial h(\boldsymbol{x}_i, \beta_n)}{\partial \beta}\frac{\partial h(\boldsymbol{x}_i, \beta_n)}{\partial \beta^\top}\left[\left(\frac{g'_{e\lambda_n}}{g_{e\lambda_n}}\right)(y_i - \boldsymbol{x}_i\beta_n) - \left(\frac{g'_0}{g_0}\right)(y_i - \boldsymbol{x}_i\beta_n)\right]^2 g_0(u)du. \tag{2.30}$$

Manski (1984) then proved that (2.30) $\rightarrow o_p(1)$ under the probability measures of the disturbances; by contiguity, it is also correct under the probability measures of the residuals.

AML estimators have been constructed, when the error distributions are assumed symmetric, for linear regression models by Bickel (1982), for nonlinear regression models by Manski (1984), for stationary ARMA processes by Kreiss (1987) and Steigerwald (1992), for linear ARCH models by Linton (1993), for cointegrating regression with ARMA errors by Hodgson (1995a) and for error correction models by Hodgson (1995c).

AML estimators are based on a first-order approximation to the likelihood function of a correctly specified model. This could cause finite sample problems simply because it is a one-step estimator and the estimator of the information matrix constructed using the OPG matrix behaves poorly in small samples. In what follows, I will consider nonlinear regression models with or without the symmetry condition and propose a new approach to constructing adaptive estimators.

3. Non-Linear Regression Models and SMLE

3.1 The Model

The model considered here is a non-linear regression model of the following form

$$y = h(\boldsymbol{x}, \boldsymbol{\beta}_0) + u, \tag{3.1}$$

where $y \in \boldsymbol{Y} \subset \boldsymbol{R}^1$, $\boldsymbol{x} \in \boldsymbol{X} \subset \boldsymbol{R}^m$ are observable, $u \in \boldsymbol{R}^1$ is unobservable. The samples $\boldsymbol{x}_i$, $i = 1, ..., n$ are *i.i.d.* with common distribution F and density function f with finite Fisher information matrix, $u_i \sim$ *i.i.d.* g_0, where f and g_0 are unknown, but g_0 is known to be absolutely continuous and symmetric around zero with respect to its argument.

With the symmetry condition on g_0, Manski (1984) proved that the parameter $\boldsymbol{\beta}_0$ satisfied the adaptive condition and was, in principle, adaptively estimable. He then proposed the adaptive maximum likelihood estimator of $\boldsymbol{\beta}_0$. In this section, we will suggest another type of semiparametric estimator which is asymptotically efficient or adaptive but is more intuitively appealing, and has well defined asymptotic properties and better finite sample performance.

3.2 The Definition of SMLE

Instead of starting from some primitive assumptions, we simply assume that the estimation problem is sufficiently regular that, for g_0, the maximum likelihood estimator of $\boldsymbol{\beta}_0$ exists. We then attempt to construct an adaptive estimator of $\boldsymbol{\beta}_0$ using the maximum likelihood approach rather than the AML approach.

The contribution to the log-likelihood by observation i is

$$\ell(y_i; \boldsymbol{x}_i, \boldsymbol{\beta}_0) \equiv \ell(y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta}_0)) = \log[g_0(y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta}_0))]. \tag{3.2}$$

When g_0 is not known, we cannot apply the ML method to estimate model (3.1). However, an initial $\sqrt{n}$-consistent estimate of $\boldsymbol{\beta}_0$ in model (3.1) can be provided by Non-Linear Least Squares or Generalized Least Squares, which is denoted as $\tilde{\boldsymbol{\beta}}_n$.

We denote the residuals $\tilde{u}_{in} = y_i - h(\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}}_n)$, $i = 1, ..., n$. These residuals will be used to construct nonparametric density functions. We choose kernel density estimation with

normal kernel in the form of (2.8) when the symmetry condition is not required; and in the form of (2.24) when the symmetry condition is required .

Consider the idealized situation where the error terms $u_1, \cdots, u_n$ in (3.1) are known. Given the bandwidth parameter λ_n, the nonparametric Kernel estimation of the density at u is $g_{e\lambda_n}(u)$. The log-density at u is then

$$\ell_{e\lambda_n}(u) = \log(g_{e\lambda_n}(u)). \tag{3.3}$$

As we discussed in Section 2, although $g_{e\lambda_n}(u)$ has well defined properties, $\ell_{e\lambda_n}(u)$, $g'_{e\lambda_n}(u)/g_{e\lambda_n}(u)$ and $g^{(2)}_{e\lambda_n}(u)/g_{e\lambda_n}(u)$ are not well behaved everywhere, where $g'_{e\lambda_n}(u)$ and $g^{(2)}_{e\lambda_n}(u)$ are as usual the first and second order derivatives of $g_{e\lambda_n}(u)$. In particular, when $g_{e\lambda_n}(u)$ is very small, we could end up with log(0), 0/0, or 1/0. To ensure good asymptotic properties of the estimators constructed based on $g_{e\lambda_n}(u)$, it is necessary to use some trimming rules to control the convergence. Relative to the AML approach, we need one more trimming parameter b_n, which is used to control the convergence of $g^{(2)}_{e\lambda_n}(u)/g_{e\lambda_n}(u)$ to $g_0^{(2)}(u)/g_0(u)$, since $g_0^{(2)}(u)/g_0(u)$ is involved in the consideration of the asymptotic efficiency of the MLE. However, it is not required by the AML approach, since the AMLE is only a first order approximation to the likelihood.

We suppress dependence on $u_1, ..., u_n$, for given $n, b_n, c_n, d_n, e_n > 0$ to trim out extreme contributions, and define the nonparametric log-density with trimming as

$\ell_{n\lambda_n}(u) =$

$$\begin{cases} \ell_{e\lambda_n}(u) & \text{if } g_{e\lambda_n}(u) \geq d_n, |u| \leq e_n, |g'_{e\lambda_n}(u)| \leq c_n g_{e\lambda_n}(u) \text{ and } |g^{(2)}_{e\lambda_n}(u)| \leq b_n g_{e\lambda_n}(u) \\ 0 & \text{otherwise} \end{cases}. \tag{3.4}$$

If u does not satisfy the four conditions in (3.4), we simply trim out its contribution to the log-likelihood, in other words, we set the weight on the contribution by u to the log-likelihood as 0. For notational purposes, we define

$g_{n\lambda_n}(u) =$

$$\begin{cases} g_{e\lambda_n}(u) & \text{if } g_{e\lambda_n}(u) \geq d_n, |u| \leq e_n, |g'_{e\lambda_n}(u)| \leq c_n g_{e\lambda_n}(u) \text{ and } |g^{(2)}_{e\lambda_n}(u)| \leq b_n g^2_{e\lambda_n}(u) \\ 1 & \text{otherwise} \end{cases}. \tag{3.5}$$

Noting that $g_{n\lambda_n}(u)$ is no longer a density function, we find, however,

$$\ell_{n\lambda_n}(u) = \log(g_{n\lambda_n}(u)). \tag{3.6}$$

This is because, when $g_{n\lambda_n}(u) = 1$, the associated log-likelihood value will be zero, which as before means we set zero weight on the log-likelihood contributed by u when u does not satisfy the four conditions in (3.4).

Since $u_1, \cdots, u_n$ in our model are unobservable, we then construct $\tilde{g}_{n\lambda_n}(u)$ and $\tilde{\ell}_{n\lambda_n}(u)$ in the same way as (3.5) and (3.6) with $\tilde{u}_{1n}$,...,$\tilde{u}_{nn}$ instead of u_1,...,u_n. We call them the pseudo nonparametric density function and the pseudo nonparametric log-density respectively. We are now ready to define the SMLE of model (3.1).

Definition 3.1.
The **SMLE** *of* β_0 *in model* (3.1) *is defined as*

$$\hat{\beta}_n \equiv \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}_{n\lambda_n}(y_i - h(\boldsymbol{x}_i, \beta)). \tag{3.7}$$

Equation (3.7) defines the SMLE of β_0, $\hat{\beta}_n$, for a nonlinear regression model. It is very intuitively appealing, since it is similar to the maximum likelihood estimator except that the log-likelihood is computed nonparametrically based on initial $\sqrt{n}$ consistent residuals. In the next two sections, we will examine the asymptotic properties of SMLE.

4. Asymptotic Properties of SMLE

To justify our subsequent analysis of the asymptotic behavior of the **SMLE**, we make some other definitions and assumptions. Following Davidson and MacKinnon (1993, Chapter 8), we define

$$\bar{\ell}(\beta; \beta_0) \equiv \text{plim}_0 \ell^n(\beta; \beta_0) \equiv \text{plim}_0 \left(n^{-1} \sum_{i=1}^{n} \ell(y_i - h(\boldsymbol{x}_i, \beta)) \right), \tag{4.1}$$

$$\bar{\ell}_{n\lambda}(\beta; \beta_0) \equiv \text{plim}_0 \ell^n_{n\lambda_n}(\beta; \beta_0) \equiv \text{plim}_0 \left(n^{-1} \sum_{i=1}^{n} \ell_{n\lambda_n}(y_i - h(\boldsymbol{x}_i, \beta)) \right), \tag{4.2}$$

where the notation "plim_0" means that the plim is calculated under the DGP characterized by (β_0, g_0). The function $\bar{\ell}(\beta, \beta_0)$ is the limiting value of n^{-1} times the log-likelihood function, and $\bar{\ell}_{n\lambda}(\beta, \beta_0)$ is the limiting value of n^{-1} times the pseudo log-likelihood function based on the true disturbance terms when the data are generated by the special case of the model with $\beta = \beta_0$ and $g = g_0$.

Assumptions.

(4.1) $\beta_0 \in B$. B *is a compact subset of* $\boldsymbol{R}^k$ *and has a nonempty interior.*

(4.2) $E \sup_{\beta \in B} |\ell(y; \boldsymbol{x}, \beta)| < +\infty$, $\ell(y; \boldsymbol{x}, \beta)$ *is a twice continuously differentiable function of* β, *and* $E[(\frac{\partial h(\boldsymbol{x}, \beta)}{\partial \beta})(\frac{\partial h(\boldsymbol{x}, \beta)}{\partial \beta^{\intercal}})]$ *is a finite positive definite matrix, given that* $|\beta - \beta_0| < \varepsilon$ *with some* $\varepsilon > 0$.

(4.3) *The asymptotic identification assumption:* β_0 *is the unique solution for the problem*

$$\max_{\beta \in B} \bar{l}(\beta; \beta_0). \tag{4.3}$$

(4.4) *For all* $g_0 \in \Phi$, $E_{\beta_0}\left[\left(\frac{g_0'}{g_0} \right)^2 (y - h(\boldsymbol{x}, \beta)) \right]$ *and* $E_{\beta_0}\left(\left| \frac{g_0^{(2)}}{g_0} \right| (y - h(\boldsymbol{x}, \beta)) \right)$ *are finite, given any* $\beta \in B$.

(4.5) g_0 *is bounded and uniformly continuous.*

The first four assumptions are regularity conditions in a strong form to ensure good asymptotic properties of the MLE, where the second condition ensures that the log-likelihood function, $\ell(y_i; \boldsymbol{x}_i, \beta)$, satisfies the uniform weak law of large numbers. The

fourth assumption ensures that $\partial \ell(y_i; \boldsymbol{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and $\partial^2 \ell(y_i; \boldsymbol{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top$ satisfy the uniform weak law of large numbers. The last assumption is not required for the asymptotic properties of the MLE, but it is required for the convergence of the sample mean of the nonparametric log-likelihood to the sample mean of the true log-likelihood, which is used to prove the uniform consistency of SMLE.

Given the contiguity of the sequence of densities of the residuals to the sequence of densities of the disturbances, we only consider the idealized situation where the true disturbances $u_1, ..., u_n$ are observed.

To show the uniform consistency of $\hat{\boldsymbol{\beta}}_n$, it suffices to prove, with the asymptotic identifiability condition, that

$$\bar{\ell}_{n\lambda}(\boldsymbol{\beta}; \boldsymbol{\beta}_0) = \bar{\ell}(\boldsymbol{\beta}; \boldsymbol{\beta}_0). \tag{4.4}$$

Using some trimming rules, we are able to control the convergence of the sample mean of the nonparametric log-likelihood to the sample mean of the true log-likelihood.

Theorem 4.1.

Under Assumptions 4.1, 4.2, 4.3, 4.4, *and* 4.5, *and given that*
$b_n \to +\infty$, $c_n \to +\infty$, $d_n \to 0$, $e_n \to \infty$,
$\lambda_n \to 0$, $n{d_n}^2{\lambda_n}^2 \to +\infty$, $e_n{\lambda_n}^5 \to o(n)$
and $n{e_n}^{-2}{\lambda_n}^2 \to +\infty$ *as* $n \to +\infty$,
then $\ell^n_{n\lambda_n}(\boldsymbol{\beta}; \boldsymbol{\beta}_0) - \ell^n(\boldsymbol{\beta}; \boldsymbol{\beta}_0) = o_p(1)$, *uniformly in* $\boldsymbol{\beta}$, *and* $\hat{\boldsymbol{\beta}}_n$ *is a uniformly consistent estimator of* $\boldsymbol{\beta}_0$.

Proof: See appendix *A*.

Theorem 4.1 proves that SMLE is consistent for models specified in Section 3.1. It is worth noting that consistency does not require the symmetry condition we made in Section 3.1. By assuming the asymptotic identification assumption, we only need to show the mean absolute convergence of $\bar{\ell}^n_{n\lambda_n}(\boldsymbol{\beta}; \boldsymbol{\beta}_0)$ to $\bar{\ell}^n(\boldsymbol{\beta}; \boldsymbol{\beta}_0)$, which can be satisfied for models with more general density functions rather than only with symmetric density functions. However, the symmetry assumption is required to provide asymptotic efficiency, where we need to examine the mean square convergence condition of the nonparametric score function.

To derive the asymptotic efficiency property, we start by considering the first-order conditions from which $\hat{\boldsymbol{\beta}}_n$ is derived. The score function for observations $i = 1, ..., n$ under g is

$$S_i(\boldsymbol{\beta}, g) \equiv -\frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{g'}{g}(y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta})), \tag{4.5}$$

and the sample mean of score vectors $S^n(\boldsymbol{\beta}, g) \equiv \frac{1}{n}\sum_{i=1}^n S_i(\boldsymbol{\beta}, g)$. The first-order conditions for $\hat{\boldsymbol{\beta}}_n$ can be written as:

$$S^n(\hat{\boldsymbol{\beta}}_n, \tilde{g}_{n\lambda_n}) = 0. \tag{4.6}$$

Then we have that

$$0 = \sqrt{n} S^n(\hat{\boldsymbol{\beta}}_n, \tilde{g}_{n\lambda_n}) = \sqrt{n} S^n(\boldsymbol{\beta}_0, \tilde{g}_{n\lambda_n})$$

$$+\frac{1}{n}\sum_{i=1}^n \frac{\partial h(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}} \frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta}^*_n)}{\partial \boldsymbol{\beta}^\top} \left\{ \frac{\tilde{g}^{(2)}_{n\lambda_n} \tilde{g}_{n\lambda_n} - [\tilde{g}'_{n\lambda_n}]^2}{[\tilde{g}_{n\lambda_n}]^2}(u^*_i) \right\} \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0), \tag{4.7}$$

where $u_i^* = y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta}_n^*)$ and $\boldsymbol{\beta}_n^*$ lies on the line segment joining $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$.

Under suitable assumptions and with specific rules for the trimming parameter b_n, we can show, without using the symmetry condition, that

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial h(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}}_n)}{\partial\boldsymbol{\beta}}\frac{\partial h(\boldsymbol{x}_i,\boldsymbol{\beta}_n^*)}{\partial\boldsymbol{\beta}^\top}\left\{\frac{\tilde{g}^{(2)}_{n\lambda_n}\tilde{g}_{n\lambda_n}-[\tilde{g}'_{n\lambda_n}]^2}{[\tilde{g}_{n\lambda_n}]^2}(u_i^*)\right\}\to -I(\boldsymbol{\beta}_0,g_0), \tag{4.8}$$

where $I(\boldsymbol{\beta}_0, g_0)$, as before, is the information matrix of $\boldsymbol{\beta}$ under $\boldsymbol{\beta}_0$ and g_0. We do not need to exploit the symmetry condition of g_0 since (4.8) can be proved by showing the mean absolute convergence condition rather than the mean square convergence condition.

Provided that $I(\boldsymbol{\beta}_0, g_0)$ is non-singular, we thus have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = (I^{-1}(\boldsymbol{\beta}_0, g_0) + o_p(1))\sqrt{n}S^n(\boldsymbol{\beta}_0, \tilde{g}_{n\lambda_n}). \tag{4.9}$$

It then suffices to show that

$$\sqrt{n}S^n(\boldsymbol{\beta}_0, \tilde{g}_{n\lambda_n}) \to N(0, I(\boldsymbol{\beta}_0, g_0)). \tag{4.10}$$

Equation (4.10) can be verified by showing the mean square convergence of $\sqrt{n}S^n(\boldsymbol{\beta}_0, \tilde{g}_{\lambda_n})$ to $\sqrt{n}S^n(\boldsymbol{\beta}_0, g_0)$. As we discussed in Section 2.3, we need to exploit the symmetry condition to prove this mean square convergence condition. Collecting the results of this section, we then have the following theorem.

Theorem 4.2:

Under Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 and given that
$b_n \to +\infty$, $c_n \to +\infty$, $d_n \to 0$,
$e_n \to \infty$, $\lambda_n \to 0$, $nd_n^2\lambda_n^2 \to \infty$.
$e_n\lambda_n{}^{-5} = o(n)$, $ne_n^{-2}\lambda_n^2 \to \infty$,
and $c_n\lambda_n \to 0$, $b_n\lambda_n \to 0$, *as* $n \to +\infty$,
then the SMLE $\hat{\boldsymbol{\beta}}_n$ *is adaptive.*

Proof: See Appendix B.

5. Non-Linear Regression Model With Free Intercept and SMLE

In Section 3, we developed adaptive estimators for nonlinear regression models under the symmetry condition, which is required, as Manski(1984) shows, for the adaptation of structural parameters in general nonlinear models. However, this fairly strict assumption is not necessary for all the sub-families of nonlinear models. One of them is a non-linear regression model with free intercept

$$y = \alpha_0 + h_1(x, \theta_0) + u. \tag{5.1}$$

Compared with model (3.1), we have

$$\beta_0 = [\alpha_0, \theta_0^\top]^\top, \quad h(x, \beta_0) = \alpha_0 + h_1(x, \theta_0), \tag{5.2}$$

where all the assumptions follow Section 3 except the symmetry of g_0, which is no longer assumed.

As we mentioned in Section 2, a weaker condition is required for the adaptation of some of the structural parameters. For model (5.1), Manski (1984) finds that θ_0 satisfies condition (2.1) when u is assumed only to be i.i.d.. θ_0 is then adaptively estimable although α_0 is not in this case. However, Manski (1984) failed to provide an adaptive estimator for θ_0, since the AMLE of θ_0 is no longer adaptive for model (5.1) without the symmetry condition. In the following, we will show that the estimators of θ_0 obtained by the procedure proposed in Section 3 are also adaptive.

Let $\hat{\alpha}_n$ and $\hat{\theta}_n$ denote respectively the estimators, with sample size n, of α_0 and θ_0 obtained from the procedure in Section 3. We have

$$S_i(\beta, g) \equiv \begin{pmatrix} S_{i\alpha}(\beta, g) \\ S_{i\theta}(\beta, g) \end{pmatrix} = \begin{pmatrix} -\frac{g'(u_i)}{g(u_i)} \\ -\frac{\partial h_1(x_i, \theta)}{\partial \theta} \frac{g'(u_i)}{g(u_i)} \end{pmatrix}. \tag{5.3}$$

Since model (5.1) is a special form of model (3.1) and the consistency of $\hat{\beta}_n$ in model (3.1) does not require the symmetry assumption, the consistency of $\hat{\alpha}_n$ and $\hat{\theta}_n$ follows:

Corollary 5.1.

Under Assumption 4.1, 4.2, 4.3 *and* 4.5, *and given that*

$b_n \to +\infty$, $c_n \to +\infty$, $d_n \to 0$, $e_n \to \infty$,

$\lambda_n \to 0$, $n{d_n}^2{\lambda_n}^2 \to +\infty$, $e_n{\lambda_n}^5 \to o(n)$

and $n{e_n}^{-2}{\lambda_n}^2 \to +\infty$ *as* $n \to +\infty$,

then $\hat{\alpha}_n$ *and* $\hat{\theta}_n$ *are uniformly consistent estimators of* α_0 *and* θ_0.

Before we examine the efficiency of $\hat{\theta}_n$, we first look at the structure of the Fisher information matrix associated with joint estimation of α_0 and θ_0 with known g_0, which is

$$I(\alpha_0, \theta_0) = \begin{pmatrix} I_{\alpha_0\alpha_0} & I_{\alpha_0\theta_0} \\ I_{\theta_0\alpha_0} & I_{\theta_0\theta_0} \end{pmatrix}$$

$$= E\left[\left(\frac{g'(u)}{g(u)}\right)^2\right] \begin{pmatrix} 1 & E\left(\frac{\partial h_1(x,\theta_0)}{\partial \theta^\top}\right) \\ E\left(\frac{\partial h_1(x,\theta_0)}{\partial \theta}\right) & E\left(\frac{\partial h_1(x,\theta_0)}{\partial \theta} \frac{\partial h_1(x,\theta_0)}{\partial \theta^\top}\right) \end{pmatrix}$$

The lower-right sub-matrix of $I^{-1}(\alpha_0, \theta_0)$ is

$$I^{\theta_0\theta_0} \equiv \left\{ E[(\frac{g'(u)}{g(u)})^2] \right\}^{-1} \left\{ E(\frac{\partial h_1(x,\theta_0)}{\partial \theta} \frac{\partial h_1(x,\theta_0)}{\partial \theta^\top}) - E(\frac{\partial h_1(x,\theta_0)}{\partial \theta}) E(\frac{\partial h_1(x,\theta_0)}{\partial \theta^\top}) \right\}^{-1}.$$

Let the score function for observation $t = 1, ..., n$ and the sample mean of score vectors under g be defined in the same way as in (4.7). We can write the first-order conditions for $\hat{\beta}_n$ as:

$$S^n(\hat{\beta}_n, \tilde{g}_{n\lambda_n}) \equiv \begin{pmatrix} S^n_\alpha(\hat{\beta}_n, \tilde{g}_{n\lambda_n}) \\ S^n_\theta(\hat{\beta}_n, \tilde{g}_{n\lambda_n}) \end{pmatrix} = 0, \tag{5.4}$$

Then we have that

$$0 = \sqrt{n}[S^n_\theta(\hat{\beta}_n, \tilde{g}_{n\lambda_n}) - E(\frac{\partial h_1(x,\theta_0)}{\partial \theta}) S^n_\alpha(\hat{\beta}_n, \tilde{g}_{n\lambda_n})]$$

$$= -\sqrt{n}\left\{ \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial h_1(x_i, \hat{\theta}_n)}{\partial \theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial \theta}) \right] \frac{\tilde{g}'_{n\lambda_n}}{\tilde{g}_{n\lambda_n}}(\hat{u}_{in}) \right\}, \tag{5.5}$$

where $\hat{u}_{in} = y_i - \hat{\alpha}_n - h_1(x_i, \hat{\theta}_n)$.

Taking a Taylor expansion of

$$\frac{g'_{n\lambda_n}}{g_{n\lambda_n}}(\hat{u}_{in}),$$

we obtain

$$0 = -\sqrt{n}\left\{ \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial h_1(x_i, \hat{\theta}_n)}{\partial \theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial \theta}) \right] \frac{\tilde{g}'_{n\lambda_n}}{\tilde{g}_{n\lambda_n}}(u_i) \right\}$$

$$+\frac{1}{n} \sum_{i=1}^n \left[\frac{\partial h_1(x_i, \hat{\theta}_n)}{\partial \theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial \theta}) \right] \frac{\partial h_1(x_i, \theta^*_n)}{\partial \theta^\top} \left\{ \frac{\tilde{g}^{(2)}_{n\lambda_n} \tilde{g}_{n\lambda_n} - [\tilde{g}'_{n\lambda_n}]^2}{[\tilde{g}_{n\lambda_n}]^2}(u^*_i) \right\} \sqrt{n}(\hat{\theta}_n - \theta_0),$$

where $u^*_i = y_i - \alpha^* - h_1(x_i, \theta^*_n)$ and $\beta^*_n = (\alpha^*_n, \theta^{*\top}_n)^\top$; as before, β^* lies on the line segment joining β_n and β_0.

Using the same argument as in Lemma B.1 in Appendix B, where the symmetry condition is not required, we can show that

$$\text{plim} \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial h_1(x_i, \hat{\theta}_n)}{\partial \theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial \theta}) \right] \frac{\partial h_1(x_i, \theta^*_n)}{\partial \theta^\top} \left\{ \frac{\tilde{g}^{(2)}_{n\lambda_n} \tilde{g}_{n\lambda_n} - [\tilde{g}'_{n\lambda_n}]^2}{[\tilde{g}_{n\lambda_n}]^2}(u^*_i) \right\}$$

$$\to -(I^{\theta_0\theta_0})^{-1}.$$

Thus, we have

$$\sqrt{n}(\theta_n - \theta_0)$$

$$= (I^{\theta_0\theta_0} + o_p(1)) - \sqrt{n}\left\{ \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial h_1(x_i, \hat{\theta}_n)}{\partial \theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial \theta}) \right] \frac{\tilde{g}'_{n\lambda_n}}{\tilde{g}_{n\lambda_n}}(u_i) \right\}. \tag{5.6}$$

It suffices to show that

$$-\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial h_1(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}-E(\frac{\partial h_1(\boldsymbol{x},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}})\right]\frac{\tilde{g}'_{n\lambda_n}}{\tilde{g}_{n\lambda_n}}(u_i)\right\}\rightarrow N(0,(I^{\boldsymbol{\theta}_0\boldsymbol{\theta}_0})^{-1}). \qquad (5.7)$$

Recall that the symmetry condition is required to show the mean square convergence of $\sqrt{n}S^n(\boldsymbol{\beta}_0,\tilde{g}_{n\lambda_n})$ to $\sqrt{n}S^n(\boldsymbol{\beta}_0,g_0)$ in Section 4. However, it is no longer required here, since the cross products in (2.24) simply drop out because

$$E\left\{\left[\frac{\partial h_1(\boldsymbol{x}_i,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}-E(\frac{\partial h_1(\boldsymbol{x},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}})\right]\left[\frac{\partial h_1(\boldsymbol{x}_j,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}-E(\frac{\partial h_1(\boldsymbol{x},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}})\right]\right\}=0,$$

when $i\neq j$.

Theorem 5.1.

Under the assumptions 4.1, 4.2, 4.3, and 4.4, and given that
$b_n\rightarrow+\infty$, $c_n\rightarrow+\infty$, $d_n\rightarrow 0$,
$e_n\rightarrow\infty$, $\lambda_n\rightarrow 0$, $nd_n^2\lambda_n^2\rightarrow\infty$,
$e_n{\lambda_n}^{-5}=o(n)$, $ne_n^{-2}\lambda_n^2\rightarrow\infty$,
and $c_n\lambda_n\rightarrow 0$, $b_n\lambda_n\rightarrow 0$, *as* $n\rightarrow+\infty$,
then the SMLE $\hat{\boldsymbol{\theta}}_n$ *is adaptive.*

Proof: See Appendix C.

6. Monte Carlo Studies

6.1. Introduction

In the above sections, we showed that it is possible, by using the SML method, to estimate nonlinear regression models asymptotically efficiently in the absence of parametric assumptions on the density of the errors. However, the foregoing description of the SML technique leaves numerous questions regarding the implementation and the finite sample performance of SML estimators. The first concern is the sensitivity of results to the selection of the smoothing and trimming parameters that appear in the expression for the kernel estimate of the density function of the innovations. The second concern is the extent of the finite sample efficiency gains by which the SMLE outperforms over the other types of estimators, especially AMLE, since part of the motivation for suggesting SMLE rather than AMLE is because we believe that SMLE has better small sample behavior than AMLE. The third concern is the performance of statistical inference based on adaptive estimators, which has not been addressed in the literature on adaptive estimation.

Little work has appeared in the literature to address the finite sample performance of adaptive estimators, with the Monte Carlo studies of Hsieh and Manski (1987), Steigerwald (1992) and Hodgson (1995b) as exceptions. Hsieh and Manski (1987) compared the finite sample performance of Adaptive Maximum Likelihood Estimators of linear regression models with various alternative parametric estimators for several different error distributions, using root mean squared error and interquartile range as the standards of comparison. They focused on the evaluation of the sensitivity of the results to smoothing and trimming parameter selection by calculating adaptive maximum likelihood estimators for a range of possible settings of these parameters.

In this section, we are concerned primarily with the finite sample performance of SMLE compared with AMLE and OLSE, rather than with how to select optimally the smoothing and trimming parameters for given samples. However, we do conduct simulations for a variety of parameter settings in order to get some idea about how sensitive finite sample efficiency gains are to the settings selected. There are also several other purposes of this study. The first is to investigate how the distance of the true distribution from the Normal distribution affects the relative efficiency gain of adaptive estimators over OLSE since, in the Normal case, all these types of estimators are asymptotically efficient. The second is to examine the effect of sample size on the relative performance of alternative estimators. Our primary criteria for evaluating performance are the root mean squared error (MSE) and the interquartile ranges of the estimators. However, to obtain evidence on the properties of hypothesis testing based on different estimation procedures, we also plot P value plots, P value discrepancy plots and size-power curves for the hypothesis tests we conduct.

6.2 The Simulated Models

The basic model used in this study is:

$$y_t = \beta_0 + \beta_1 x_t + u_t, \tag{6.1}$$

$$(\beta_0, \beta_1) = (1.0, 1.0), \tag{6.2}$$

where x_t is *i.i.d.* and drawn from standard normal; u_t is *i.i.d.*, independent of x_t and drawn from $g_0(u)$, which takes on each of the following forms (all normalized to have zero mean and unit variance):

($a1$) t with three degrees of freedom (t3);
($a2$) t with five degrees of freedom (t5);
($a3$) t with eight degrees of freedom (t8);
($a4$) standard normal (N);
($b1$) bimodal symmetric mixture of two normals, $0.5(N(-3,1) + N(3,1))$ (BN1);
($b2$) bimodal symmetric mixture of two normals, $0.5(N(-2,1) + N(2,1))$ (BN2);
($b3$) bimodal symmetric mixture of two normals, $0.5(N(-1,1) + N(1,1))$ (BN3);
($c1$) log normal, $\log(u)$ is $N(0,1)$ (LN1).
($c2$) log normal, $\log(u)$ is $N(0,0.5)$ (LN2).
($c3$) log normal, $\log(u)$ is $N(0,0.1)$ (LN3).

In order to identify the intercept parameter β_0, we need to center, where necessary, the various distributions at mean zero. For the purposes of comparability, we also rescale the above distributions so that they all have a standard deviation equal to one.

The distributions in Group a are normalized student-t distributions with different degrees of freedom. They are *symmetric* but *leptokurtic* and differ from a normal density in that they have higher tail probabilities. When the degrees of freedom go to infinity, the t distribution goes to the Normal distribution. The distributions in Group b are normalized bimodal symmetric mixtures of normal distributions. They are also *symmetric* and *leptokurtic*. When the mean parameter in the normal distributions goes to zero. the normalized bimodal symmetric mixture of normals goes to the normal distribution. The distributions in Group c are normalized log-normals with different variance from the normal. They are *asymmetric*. When the variance of the normal goes to zero, the normalized log-normal distribution goes to the normal distribution.

When $g_0(u)$ takes on any of the forms in Group a and b, both SML and AML estimators for β_0 and β_1 are adaptive. They should theoretically outperform OLSE. When $g_0(u)$ takes on any of the forms in Group c, only SMLE for β_1 is adaptive. SMLE for β_0 in these cases is consistent.

6.3 The Relative Efficiency

In this section, we study the relative efficiency of alternative estimators. Our primary criteria for evaluating efficiency are the root mean squared error (MSE) and the interquartile ranges of the estimators, which indicate the precision of the estimation.

Implementation of the Study

As we mentioned before, one concern which appears in this Monte Carlo study is how to select the smoothing and trimming parameters called for in the nonparametric density estimator. The second is how to compare the performance of alternative estimators.

Asymptotic theory requires only that the amounts of smoothing and trimming imposed be reduced or increased at appropriate rates as the sample size increases. Given a fixed sample size, these restrictions on rates in no way constrain our choice of magnitudes for the smoothing and trimming parameters. Hsieh and Manski (1987) examined the finite sample performance of AMLE via Monte Carlo methods. They focused on the problem of selecting the smoothing and trimming parameters used in estimating the score function. They compared the AMLE when these parameters were preselected or, alternatively, were determined by a data-based bootstrap method.

The trimming parameters have been used in the AML procedure to control the mean square convergence of the nonparametric estimation of the score function to the true score function g_0'/g_0. They are introduced in the following way.

$$S_{n\lambda_n}(u) = \begin{cases} \frac{g'_{e\lambda_n}}{g_{e\lambda_n}}(u) & \text{if } |u| \le e_n, g_{e\lambda_n}(u) \ge d_n, \text{ and } |\frac{g'_{e\lambda_n}}{g_{e\lambda_n}}(u)| \le c_n, \\ 0 & \text{otherwise,} \end{cases}$$

where $g_{e\lambda_n}(u)$ is the nonparametric density function estimate of g_0. $g_{e\lambda_n}(u)$ takes the form of (2.24) when g_0 is symmetric; otherwise, it takes the form of (2.8). In the SML procedure, we introduced the trimming parameters in the following way:

$\ell_{n\lambda_n}(u) =$

$$\begin{cases} \ell_{e\lambda_n}(u) & \text{if } g_{e\lambda_n}(u) \ge d_n, |u| \le e_n, |g'_{e\lambda_n}(u)| \le c_n g_{e\lambda_n}(u) \text{ and } |g^{(2)}_{e\lambda_n}(u)| \le b_n g_{e\lambda_n}(u) \\ 0 & \text{otherwise} \end{cases},$$

where we trim out the log-likelihood contributed by an observation which does not satisfy the four conditions described in the above equation. Note that, in the SML procedure, we need the trimming parameter b_n; while b_n is not required by the AML procedure.

For sample size 50. Hsieh and Manski (1987) selected λ from a number of values ranging from 0.01 to 2.00 and $[e_n = \rho, d_n = \exp(-\rho^2/2), c_n = \rho]$, where ρ was allowed to take on the values of 3, 4 and 8. They found that the best performances of AMLE happened when ρ was 8 and λ was between 0.05 and 0.5 varying across distributions.

Since how to select these parameters optimally is not our main concern, we simply follow Hsieh and Manski (1987) by choosing $[e_n = \rho, d_n = \exp(-\rho^2/2), c_n = \rho]$, however, with sample size 100 we allow ρ to take the values 8 and 12, and λ to take the values 0.40 and 0.50 to get some idea of how these parameters affect the estimates. We set $b_n = \infty$ for purposes of comparability simply because there is no estimate involved for the second order derivative of density function for AMLE. We use the normal density as a kernel to estimate the nonparametric density function. In each case, 1000 replications were done.

In previous Monte Carlo studies, the authors tabulated only root MSE and interquartile ranges for alternative estimators. We argue that these tables cannot provide all the important information obtained from the Monte Carlo studies. We there plot some of the empirical distributions of alternative estimators. From the empirical distributions of alternative estimators, we can see at a glance why alternative estimators behave differently.

The Results with Sample Size 100

Tables 1-10 report results on MSE and interquartile ranges for the SML, AML and OLS estimators.

The findings can be summarized below.

Comparability of root mean square error and quartile results. Consistent with the results in Hsieh and Manski (1987), measurements of the precision of an estimator by root mean square error and by interquartile range yield very similar results, in the sense that these two statistics almost always rank the various estimators in the same order.

Effect of trimming and smoothing parameters. Given the trimming and smoothing parameters that have been chosen, we find the results are fairly insensitive to these selections.

Performance of SMLE relative to OLSE. We calculate the ratio of the root MSE of SML estimators to the root MSE of OLS estimators only for one set of trimming and smoothing parameter selection, $\rho = 12$ and $\lambda = 0.5$, since the estimation results are fairly insensitive to the selection. The results are presented in Table 11. In the cases of symmetric densities, student-t and bimodal symmetric mixture of two normals, we find SMLE outperforms the OLSE for both the parameters β_0 and β_1 except the normal case. The relative efficiency gains of SMLE over OLSE range from 2 to 64 per cent.

In the log-normal cases, SMLE outperforms OLSE only for the parameter β_1, not β_0. This is consistent with our theoretical findings that the SML method only provides adaptive estimator of β_1 rather than both β_0 and β_1 when the error distribution is asymmetric.

Not surprisingly, we find that the more the distribution differs from the normal distribution, the greater are the efficiency gains achieved by SMLE over OLSE. In the student-t cases, the degrees of freedom indicate the departure of the t-distribution from the normal distribution. Smaller degrees of freedom indicate greater departure of the t-distribution from the normal distribution. When the true distribution is normal, which is a t-distribution with infinite degrees of freedom, the relative efficiency gains of SMLE over OLSE vanish and OLSE slightly outperforms SMLE. In the bimodal symmetric mixture of two normals cases, the mean of the normals indicates the departure of these distributions from the normal distribution. Greater mean represents greater departure. We find that when the mean goes to zero, the relative efficiency gains of SMLE over OLSE vanish. Similar results can also be found for the estimation of β_1 in the log-normal cases. When the variance goes to zero, the relative efficiency gains of SMLE over OLSE diminish.

Performance of AMLE relative to OLSE. Table 13 presents the ratio of the root MSE of AMLE to the root MSE of OLSE. We find that AMLE outperforms OLSE for both the parameters β_0 and β_1 in the cases of symmetric densities except the normal and BN3 cases. The more distance the true density differs from normal density, the greater the relative efficiency gains AMLE achieves over OLSE. However, in the log-normal cases, AMLE performs worse than OLSE, especially for the estimation of β_0. This is because AMLE is no longer adaptive for the asymmetric densities.

Performance of SMLE relative to AMLE. The ratio of the root MSE of SMLE to the root MSE of AMLE is calculated and reported in Table 12. SMLE almost always provides substantial efficiency gains over AMLE. The efficiency gains achieved by SMLE over AMLE range from 2 to 70 per cent.

In Figures 1-3, we plot the empirical distribution functions of the estimates of β_1 for the t3, BN1 and LN1 cases. These plots more clearly illustrate the ranking of alternative estimators. In the cases of symmetric densities, SMLE performs the best, and OLSE the worst. However, in the asymmetric cases, SMLE performs the best, and AMLE the worst.

The consistency of SMLE and AMLE in the log-normal cases. As we showed in Section 4, the consistency of SMLE requires no symmetry assumption for the error distributions. We tabulate in Table 14 the average of the estimates for β_0 in the log-normal cases when $\rho = 12$ and $\lambda = 0.5$.

The Effect of the Sample Size

To examine the effect of the sample size on relative performance, we also conduct Monte Carlo experiments for the model (6.1) and (6.2) with sample sizes 50 and 200. The distributions considered are respectively t3, BN1 and LN1; ρ and λ are set to be 12 and 0.5. Since the measurements of the precision of an estimator by root mean square error and by interquartile range yield very similar results, we only report the root mean square error for these experiments. Table 15 reports the root MS estimators of alternative estimators. The results for relative efficiency, which is measured by the ratio of the root MS estimators, are presented in Table 16. We find that, in the symmetric cases, SML estimators of both β_0 and β_1 consistently perform the best and OLS estimators perform the worst. In the asymmetric case, SMLE outperforms OLSE for the parameter β_1 and AMLE consistently performs the worst for both β_0 and β_1.

6.4 Inference Based on Different Estimation Methods

In addition to relative efficiency, another concern for an estimation method is whether we can draw correct inferences from this method. This has not been addressed by the aforementioned authors although it is as important as relative efficiency. In our Monte Carlo studies, we will compare the finite-sample properties of the t-ratio test obtained from alternative estimation procedures. The t-ratio values are computed for the following tests.

Test 1
the null hypothesis: $\beta_1 = 1$,
the alternative: $\beta_1 \neq 1$.
Test 2
the null hypothesis: $\beta_1 = 0.9$,
the alternative: $\beta_1 \neq 0.9$.

The data is generated from the DGP described in (6.1) and (6.2), where $\beta_1 = 1$. Therefore, the null hypothesis is true in the first test, however, the null hypothesis is false in the second test.

In calculating the t-ratio statistic for SMLE and AMLE, we estimate the covariance matrices of SMLE and AMLE using the form of estimation $H^{-1}(G^{\top}G)H^{-1}$, where H and G are respectively the sample Hessian matrix and the matrix of contributions to the gradient. We use this form of estimation in the consideration of robustness. To present our results, we use the graphical method developed by Davidson and MacKinnon (1994), which is the best way we find to report Monte Carlo results of hypothesis testing procedures.

The P value plots and P value discrepancy plots

In each Monte Carlo experiment, 1000 realizations of the t-statistic τ are generated by each estimation method. We denote these simulated values by τ_j, $j = 1, \cdots, 1000$. The P value of τ_j is the probability of observing a value of τ as or more extreme than τ_j, according to some distribution of τ, which in our case is the asymptotic distribution of τ, i.e. the standard normal distribution (two tails). The P value associated with τ_j is denoted as $p_j \equiv p(\tau_j)$. The empirical distribution function of the p_j's is simply an estimate of the c.d.f. of $p(\tau)$. At any point z_k in the $(0, 1)$ interval, it is defined by

$$\hat{F}(z_k) \equiv \frac{1}{1000} \sum_{j=1}^{1000} I(p_j \leq z_k) \tag{6.3}$$

where $I(p_j \leq z_k)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The simplest graph is a plot of $\hat{F}(z_k)$ against z_j, which is called a P **value plot**. If the distribution of τ used to compute the p_j's is correct, each of the p_j's should be close to the 45° line. P value plots allow us to distinguish at a glance among test statistics that systematically over-reject, test statistics that systematically under-reject, and test statistics that reject about the right proportion of the time. However, to distinguish well-behaved test statistics, it is much more revealing to graph $\hat{F}(z_k) - z_k$, which is called a P **value discrepancy plot**. We use P value plots and P value discrepancy plots to report the results of the first test from the experiments conducted .

P value plots and P value discrepancy plots. Figures 4-6 show P value plots for the t-ratio tests from different estimation methods in the $t3$, $BN1$ and $LN1$ cases when the sample size $n = 100$. Figures 7-9 show the associated P value discrepancy plots. These figures make it clear that the t-ratio test of OLSE works the best, and the t-ratio test of AMLE performs the worst, especially for the log-normal case. This is not surprising given the small sample size, since the covariance matrix estimates of SMLE and AMLE are quite noisy. Figures 10-12 show P value plots for the t-ratio tests from the different estimation methods in the t3, BN1 and LN1 cases when the sample size $n = 200$. Figures 13-15 show the associated P value discrepancy plots. We find that when the sample size $n = 200$, the t-ratio tests computed from the SMLE procedure and the OLSE procedure work equally well, however, the t-ratio test computed from the AMLE procedure remains the worst.

The size-power curves

In the second test, the null hypothesis is false. It is not very useful to plot the EDF's of P values of the calculated test statistics, since we will be plotting power against *nominal* test size. In order to obtain informative results, we will plot the size-power curve proposed by Davidson and MacKinnon (1993, Chapter 12) and Davidson and MacKinnon (1994). To do so, we perform two experiments using the same sequence of random numbers. In the first experiment, we set $\beta_1 = 0.9$, which means the null hypothesis holds. In the second, $\beta_1 = 1$, where the null hypothesis is false. Let the points on the two approximate EDF's be denoted $\hat{F}(z)$ and $\hat{F}^*(z)$. They are evaluated as before at a prechosen set of points z_k, $k = 1, \cdots, m$. The size-power curve is obtained by plotting $\hat{F}^*(z_k)$ against $\hat{F}(z_k)$. For sample size $n = 100$, we calculate $\hat{F}(z_k)$ and $\hat{F}^*(z_k)$ in the cases of t3, BN1, and LN1. We present the size-power curves in Figures 16-18. It is clear that for all these cases, the t-ratio test computed from the SML procedure has the greatest power for a given size. In contrast, the t-ratio test computed from the OLS procedure has the least power for a given size in the symmetric cases and the t-ratio test computed from the AML procedure has the least power for a given size in the asymmetric case.

7. Conclusion

We have developed semiparametric maximum likelihood estimators for the structural parameters in a general nonlinear regression model with or without the symmetry condition. Using SML estimation method, we have constructed adaptive estimators not only for all the structural parameters in a nonlinear regression model with the symmetry condition but also for the adaptively estimable parameters in a nonlinear regression model with a free intercept without the symmetry condition. We have also found that SML estimators are generally consistent for all the structural parameters with or without the symmetry condition.

Part of the motivation for suggesting SMLE rather than AMLE is that we expect better small sample performance of SMLE over AMLE. We have conducted Monte Carlo studies to examine the small sample performance of SMLE compared to AMLE and OLSE. Based on the MSE and interquartile ranges of the estimators, we find that SMLE performs the best and OLSE the worst for all the structural parameters when the true density is symmetric. The greater the distance of the true density function from the normal density, the greater the efficiency gains SMLE and AMLE achieve over OLSE. In the asymmetric density cases, SMLE performs the best and AMLE the worst for the adaptively estimable parameter; the distance of the true density from the normal density also has a similar result on the relative efficiency gains of SMLE over OLSE.

In this Monte Carlo study, we also study the relative performance of alternative estimators with different sample sizes. We find that SMLE consistently performs the best for the adaptively estimable parameters.

Using P value plots and P value discrepancy plots, we show that the t-ratio test computed from the AML procedure tends to over-reject the null hypothesis, especially in the left tail. In contrast, the t-ratio test computed from the OLS procedure works quite well. And the t-ratio test computed from the SML procedure works reasonably well with the sample size $n = 100$; it works quite well when the sample size $n = 200$.

Using size-power curves, we show that the t-ratio tests computed from the SML procedure has the greatest power for a given size of test. In contrast, the t-ratio test computed from the OLSE procedure has the least power.

Literature

Begun, J., W. Hall, W. Huang and J. Wellner (1983), "Information and Asymptotic Efficiency in Parametric-Nonparametric Models", *Annals of Statistics*, 11, 432-452.

Beran, R. (1974), "Asymptotically Efficient Adaptive Rank Estimates in Location Models", *Annals of Statistics*, 2, 63-74.

Bickel, P. (1982), "On Adaptive Estimation", *Annals of Statistics*, 10, 647-671.

Bollerslev, T. P. (1987), "A Conditional Heteroscedastic Time Series Model for Security Prices and Rates of Return Data," *Review of Economics and Statistics*, 69, 542-547.

Davidson, R. and J. G. MacKinnon (1993), "Estimation and Inference in Econometrics", Oxford University Press, New York.

Davidson, R. and J. G. MacKinnon (1994), "Graphical methods for investigating the size and power of hypothesis tests", *The Manchester School*, forthcoming.

Devroye, L. and L. Gyorfi (1985), "Nonparametric Density Estimation, THE L_1 VIEW", John Wiley & Sons.

Engle, R. F. and G. González-Rivera (1991), "Semiparametric ARCH Models," *Journal of Business & Economic Statistics*, Vol.9, No.4, 345-359.

Hajek, J. and Z. Sidak (1967), "Theory of Rank Tests", Academic Press, New York.

Hodgson, D. (1995a), "Adaptive Estimation of Cointegrating Regressions with ARMA Errors", Working Paper No. 408, Rochester Center for Economic Research.

Hodgson, D. (1995b), "Adaptive Estimation of Cointegrated Models Simulation Evidence and an Application to the Forward Exchange Market", Working Paper No. 409, Rochester Center for Economic Research.

Hodgson, D. (1995c), "Adaptive Estimation of Error Correction Models", Working Paper No. 410, Rochester Center for Economic Research.

Hsieh, D. and C. Manski (1987), "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression", *Annals of Statistics*, Vol. 15, No. 2, 541-551.

Kreiss, J. (1987), "On Adaptive Estimation in Stationary ARMA Processes", *Annals of Statistics*, 15(1), 112-133.

LeCam, L. (1953), "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes's Estimates", University of California Publications in Statistica, 1, 277-330.

LeCam, L. (1960), "Locally Asymptotically Normal Families of Distributions", University of California Publications in Statistics, 3, 37-98.

Linton, O. (1993), "Adaptive Estimation in ARCH Models", *Econometric Theory*, 9, 539-569.

Manski, C. (1984), "Adaptive Estimation of Non-linear Regression Models", *Econometric Reviews*, 3(2), 145-194.

Newey W. and D. Steigerwald (1994), "Consistency of Quasi-Maximum Likelihood Estimators for Models with Conditional Heteroscedasticity", University of California, Santa Barbara, Working Paper in Economics, #5-94.

Peruga, R. (1988), "The Distributional Properties of Exchange Rate Changes Under a Peso Problem," unpublished Ph.D. dissertation, University of California, San Diego, Dept. of Economics.

Prakasa Rao, B. L. S. (1983), "Nonparametric Functional Estimation", Academic Press.

Steigerwald, D. (1992), "Adaptive Estimation in Time Series Regression Models", *Journal of Econometrics*, 54, 251-275.

Stein, C. (1956), "Efficient Nonparametric Testing and Estimation", Proc. Third Berkeley Symp. Math. Statist. Prob., 1, 187-196, University of California Press, Berkeley.

Stone, C. (1975), "Adaptive Maximum Likelihood Estimation of a Location Parameter", *Annals of Statistics*, 3, 267-284.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1-26.

Appendix A

We define the following metrics used in the appendix to prove our theorems. $\boldsymbol{L}^1(z) = E(|z|)$, $\boldsymbol{L}^2(z) = [E(z^2)]^{1/2}$. The relationship between them is

$$\boldsymbol{L}^1(z) \leq \boldsymbol{L}^2(z), \tag{A.1}$$

where z is a random variable.

Proof of Theorem 4.1:

We begin by recalling the following Lemma.

Lemma *A.1.*

Let g be bounded and uniformly continuous. Then there exist positive constants c_1 and c_2 such that for any sequence of positive numbers, $\{\varepsilon_n\}$,

$$P\{\sup_u |g(u) - g_{e\lambda_n}(u)| > \varepsilon_n\} \leq c_1 \exp(-c_2 n {\varepsilon_n}^2 {\lambda_n}^2), \tag{A.2}$$

whenever $\lambda_n = o(\varepsilon_n)$.

Proof: See Prakasa Rao (1983), Section 4.1.

To prove Theorem 4.1, we use L^1 metric and let

$$I \equiv \int \cdots \int \int \cdots \int n^{-1} \sum_{i=1}^{n} \left| \ell_{n\lambda_n}(y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta})) - \ell(y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta})) \right|$$

$$g_0[y_1 - h(\boldsymbol{x}_1, \boldsymbol{\beta})] \cdots g_0[y_n - h(\boldsymbol{x}_n, \boldsymbol{\beta})] f(\boldsymbol{x}_1) \cdots f(\boldsymbol{x}_n) dy_1 \cdots dy_n d\boldsymbol{x}_1 \cdots d\boldsymbol{x}_n. \qquad (A.3)$$

Then, we have, by i.i.d. property,

$$I = \int \int \left| \ell_{n\lambda_n}(y - h(\boldsymbol{x}, \boldsymbol{\beta})) - \ell(y - h(\boldsymbol{x}, \boldsymbol{\beta})) \right| g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})] f(\boldsymbol{x}) dy d\boldsymbol{x}. \qquad (A.4)$$

Denote the conditions in (3.4) by A, B, C, D, and the intersection regions in $(y, \boldsymbol{x})$ space as $ABCD$. The right hand side of $(A.4)$ is denoted by $I_1 + I_2$, where

$$I_1 = \int \int_{ABCD} \left| \log(g_{e\lambda_n}[y - h(\boldsymbol{x}, \boldsymbol{\beta})]) - \log(g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})]) \right| g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})] f(\boldsymbol{x}) dy d\boldsymbol{x},$$

$$I_2 = \int \int_{[ABCD]^c} \left| \log[g_0(y - h(\boldsymbol{x}, \boldsymbol{\beta}))] \right| g_0(y - h(\boldsymbol{x}, \boldsymbol{\beta})) f(\boldsymbol{x}) dy d\boldsymbol{x}.$$

Because $|\log(z)| \leq \left| \frac{1}{z} - 1 \right| + \left| z - 1 \right|$ for all scalar $z > 0$ (see Prakasa Rao, 1983, Section 4.4), then I_1 is bounded by

$$\int \int_{ABCD} \left\{ \left| \frac{g_{e\lambda_n}[y - h(\boldsymbol{x}, \boldsymbol{\beta})] - g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})]}{g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})]} \right| \right.$$

$$\left. + \left| \frac{g_{e\lambda_n}[y - h(\boldsymbol{x}, \boldsymbol{\beta})] - g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})]}{g_{e\lambda_n}[y - h(\boldsymbol{x}, \boldsymbol{\beta})]} \right| \right\} g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})] f(\boldsymbol{x}) dy d\boldsymbol{x}. \qquad (A.5)$$

Recall that $g_{e\lambda_n}[y - h(\boldsymbol{x}, \boldsymbol{\beta})] \geq d_n$ and $y - h(\boldsymbol{x}, \boldsymbol{\beta}) \leq e_n$, $(A.5)$ is then bounded by

$$\sup_u |g_{e\lambda_n}(u) - g_0(u)| \int \int_{ABCD} \left\{ \frac{1}{g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})]} + \frac{1}{d_n} \right\} g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})] f(\boldsymbol{x}) dy d\boldsymbol{x}$$

$$\leq \sup_u |g_{e\lambda_n}(u) - g_0(u)| (2e_n + \frac{1}{d_n}). \qquad (A.6)$$

Given the conditions $nd_n^2\lambda_n^2 \to \infty$ and $ne_n^{-2}\lambda_n^2 \to \infty$, then $I_1 \to o_p(1)$ uniformly in $\boldsymbol{\beta}$ by Lemma $A.1$.

We bound I_2 by

$$I_2 \leq \int \int \left| \log(g_0[y - h(\boldsymbol{x}, \boldsymbol{\beta})]) \right| [P\{|g^{(2)}_{e\lambda_n}[y - f(\boldsymbol{x}, \boldsymbol{\beta})]| > b_n g_{e\lambda_n}[y - f(\boldsymbol{x}, \boldsymbol{\beta})]\}$$

$$+P\{|g'_{e\lambda_n}[y-f(\boldsymbol{x},\boldsymbol{\beta})]| > c_n g_{e\lambda_n}[y-f(\boldsymbol{x},\boldsymbol{\beta})]\} \tag{A.7}$$

$$+P\{g_{e\lambda_n}[y-f(\boldsymbol{x},\boldsymbol{\beta})] < d_n\} + P\{|y-h(\boldsymbol{x},\boldsymbol{\beta})| > e_n\}]g_0(y-h(\boldsymbol{x},\boldsymbol{\beta}))f(\boldsymbol{x})dyd\boldsymbol{x}.$$

In order to show that the left hand side of $(A.7)$ tends to 0 in probability, it suffices to prove that the four probabilities at the left hand side of $(A.7)$ tends to 0.

Using the elementary estimates noted in Stone (1975), with some constants κ_i for all g_0 and all u, we then have

$$\text{Var } g^{(i)}_{e\lambda_n}(u) \leq \kappa_i {\lambda_n}^{-(2i+1)} n^{-1} g_{\lambda_n}(u),\ i = 0, 1, ... \tag{A.8}$$

Note that $g_{\lambda_n}(u)$ is the convolution function of g_0 and ϕ_{λ_n}; the variance in $(A.8)$ is an operator on $u_1, \cdots, u_n$.

We claim that

$$g_{e\lambda_n}(u) \to g_0(u) \text{ in probability for all } u \text{ if } n\lambda_n \to +\infty, \tag{A.9}$$

$$g'_{e\lambda_n}(u) \to g'_0(u) \text{ in probability } a.e.\ u \text{ if } n{\lambda_n}^3 \to +\infty, \tag{A.10}$$

$$g^{(2)}_{e\lambda_n}(u) \to g^{(2)}_0(u) \text{ in probability } a.e.\ u \text{ if } n{\lambda_n}^5 \to +\infty. \tag{A.11}$$

where $a.e.$ means almost everywhere.

By the assumptions 4.1-4.4, $P\{|g^{(2)}_0[y-f(\boldsymbol{x},\boldsymbol{\beta})]| > b_n g_0[y-f(\boldsymbol{x},\boldsymbol{\beta})]\}$, $P\{|g'_0[y-f(\boldsymbol{x},\boldsymbol{\beta})]| > c_n g_0[y-f(\boldsymbol{x},\boldsymbol{\beta})]\}$, $P\{g_0[y-f(\boldsymbol{x},\boldsymbol{\beta})] < d_n\}$, and $P\{|y-h(\boldsymbol{x},\boldsymbol{\beta})| > e_n\}$ all tend to 0 when $n \to \infty$, then $I_2 \to o_p(1)$. The Theorem is proved.

It now remains to prove $(A.9) - (A.11)$. By $(A.8)$, for all u, we have

$$g_{e\lambda_n}(u) \to g_{\lambda_n}(u) \text{ in probability if } n\lambda_n \to +\infty, \tag{A.12}$$

$$g'_{e\lambda_n}(u) \to g'_{\lambda_n}(u) \text{ in probability if } n{\lambda_n}^3 \to +\infty. \tag{A.13}$$

$$g^{(2)}_{e\lambda_n}(u) \to g^{(2)}_{\lambda_n}(u) \text{ in probability if } n{\lambda_n}^5 \to +\infty. \tag{A.14}$$

Continuity of g_0 and $(A.12)$ implies $(A.9)$. To prove $(A.10)$, we write

$$\int |g'_{\lambda_n}(u) - g'_0(u)|du = \int \left| \int \left[g'_0(u-\lambda_n z) - g'_0(u)\right] \phi(z) dz \right| du$$

$$\leq \int \int \left|g'_0(u-\lambda_n z) - g'_0(u)\right| du\ \phi(z)\ dz. \tag{A.15}$$

where $\phi(z)$ is the density function of standard normal. Note that $\int |g'_0(u)|du < \infty$, which is implied by Assumption 4.4. We thus can apply the $\boldsymbol{L}_1$ continuity theorem and the dominated convergence theorem to conclude that the right-hand side of $(A.15)$ tends to 0 as $\lambda_n \to 0$ and $(A.10)$ follows from $(A.13)$ and $(A.15)$. Using the same argument, we can also prove $(A.11)$ by $(A.14)$ and Assumption 1.4.4.

Appendix B

Proof of Theorem 4.2:
We begin by proving the following three Lemmas.
Lemma $B.1$:
Let

$$M(\hat{\beta}_n, \beta_n^*, g_0) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(x_i, \hat{\beta}_n)}{\partial \beta} \frac{\partial h(x_i, \beta_n^*)}{\partial \beta^\top} \left(\frac{g_0^{(2)} g_0 - [g_0']^2}{[g_0]^2} \right)(u_i^*) \qquad (B.1)$$

$$M(\hat{\beta}_n, \beta_n^*, g_{n\lambda_n}) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(x_i, \hat{\beta}_n)}{\partial \beta} \frac{\partial h(x_i, \beta_n^*)}{\partial \beta^\top} \left(\frac{g_{n\lambda_n}^{(2)} g_{n\lambda_n} - [g_{n\lambda_n}']^2}{[g_{n\lambda_n}]^2} \right)(u_i^*). \qquad (B.2)$$

By that both $\hat{\beta}_n$ *and* β_n^* *tend to* β_0, *when* $n \to \infty$, *then*

$$\operatorname{plim} M(\hat{\beta}_n, \beta_n^*, g_{n\lambda_n}) = \operatorname{plim} M(\hat{\beta}_n, \beta_n^*, g_0) \to -I(\beta_0, g_0), \ \textit{when } \lambda_n \to 0. \qquad (B.3)$$

Proof: We use $\boldsymbol{L}^1$ metric to prove this Lemma. Let

$$J = \int \cdots \int \int \cdots \int |M(\hat{\beta}_n, \beta_n^*, g_{n\lambda_n}) - M(\hat{\beta}_n, \beta_n^*, g_0)|$$

$$g_0(u_1) \cdots g_0(u_n) f(x_1) \cdots f(x_n) du_1 \cdots du_n dx_1 \cdots dx_n.$$

With i.i.d. $u_1, ..., u_n$, and i.i.d. $x_1, ..., x_n$, we have

$$J = \int \int \left| \frac{\partial h(x, \beta_0)}{\partial \beta} \frac{\partial h(x, \beta_0)}{\partial \beta^\top} + o_p(1) \right|$$

$$\left| \left[\frac{g_{n\lambda_n}^{(2)}}{g_{n\lambda_n}} - \frac{g_0^{(2)}}{g_0} \right](u_1) - \left[\frac{[g_{n\lambda_n}']^2}{[\tilde{g}_{n\lambda_n}]^2} - \frac{[g_0']^2}{[g_0]^2} \right](u_1) + o_p(1) \right| g_0(u_1) f(x)\, du_1\, dx, \qquad (B.4)$$

where $g_{n\lambda_n}(u_1)$ is the pseudo nonparametric density estimate of $g_0(u_1)$ given $u_2, \cdots, u_n$. For notational simplicity, we will, in the following calculation, use u instead of u_1 in $(B.4)$. Then $g_{n\lambda_n}(u)$ can be treated asymptotically as the pesudo nonparametric density estimate of $g_0(u)$ given $u_1, \cdots, u_n$. We suppress the dependence on u in the following four results.

$$\int \left| \left[\frac{g_{n\lambda_n}^{(2)}}{g_{n\lambda_n}} - \frac{g_0^{(2)}}{g_0} \right] - \left[\frac{[g_{n\lambda_n}']^2}{[g_{n\lambda_n}]^2} - \frac{[g_0']^2}{[g_0]^2} \right] \right| g_0 du \qquad (B.5)$$

can be shown to be $o_p(1)$ by the following four results.
Result 1:

$$\int \left| \frac{[g_{n\lambda_n}']^2}{[g_{n\lambda_n}]^2} - \frac{[g_0']^2}{[g_0]^2} \right| g_0 du \to o_p(1).$$

Proof: Bickel (1982), Section 6 proved the mean square convergence of $g'_{n\lambda_n}/g_{n\lambda_n}$ to g'_0/g_0. Given that the conditions he assumed on the trimming and smooth parameters are all satisfied, Result 1 simply follows.

In order to show that $(B.5)$ is $o_p(1)$, it remains to prove that

$$\int \left| \frac{g^{(2)}_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g^{(2)}_0}{g_0} \right| g_0 du \to o_p(1). \tag{B.6}$$

The left hand side of $(B.6)$ can be bounded by

$$\int \left| \frac{g^{(2)}_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g^{(2)}_{\lambda_n}}{g_{\lambda_n}} \right| g_0 du \leq \int \left| \frac{g^{(2)}_{n\lambda_n}}{g_{n\lambda_n}} (g_0 - g_{\lambda_n}) \right| du$$

$$+ \int \left| \frac{g^{(2)}_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g^{(2)}_{\lambda_n}}{g_{\lambda_n}} \right| g_{\lambda_n} du + \int |g^{(2)}_{\lambda_n} - g^{(2)}_0| du. \tag{B.7}$$

The three terms are shown to be $o_p(1)$ by respectively **Result 3**, **Result 2**, and **Result 4** as follows.

Result 2:

Let

$$I_3 = \int \left| \frac{g^{(2)}_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g^{(2)}_{\lambda_n}}{g_{\lambda_n}} \right| g_{\lambda_n} du, \tag{B.8}$$

given that $b_n\lambda_n \to 0$ and that $e_n\lambda_n^{-5} = o(n)$, we then have $I_3 \to o_p(1)$.

Proof:

$$I_3 = \int_{[ABCD]} \left| \frac{g^{(2)}_{e\lambda_n}}{g_{e\lambda_n}} - \frac{g^{(2)}_{\lambda_n}}{g_{\lambda_n}} \right| g_{\lambda_n} du + \int_{[ABCD]^c} \left| \frac{g^{(2)}_{\lambda_n}}{g_{\lambda_n}} \right| g_{\lambda_n} du = I_4 + I_5. \tag{B.9}$$

In order to prove that $I_5 \to o_p(1)$. we first present the result that $\int |g^{(2)}_{\lambda_n}| du \leq \int |g^{(2)}_0| du$. This is because

$$\int |g^{(2)}_{\lambda_n}(u)| du = \int \left| \int g^{(2)}_0(u-z) \phi_{\lambda_n}(z) dz \right| du$$

$$\leq \int \left\{ \int |g^{(2)}_0(u-z)| \phi_{\lambda_n}(z)\, dz \right\} du$$

$$= \int \left\{ \int |g^{(2)}_0(u-z)| du \right\} \phi_{\lambda_n}(z)\, dz$$

$$= \int \phi_{\lambda_n}(z) dz.$$

We bound $E(I_5)$ by

$$E(I_5) \leq \int \cdots \int \int |g^{(2)}_{\lambda_n}| \left[P\{|g^{(2)}_{e\lambda_n}| > b_n g_{e\lambda_n}\} + P\{|g'_{e\lambda_n}| > c_n g_{e\lambda_n}\} \right.$$

$$+P\{g_{e\lambda_n} < d_n\} + P\{|u| > e_n\}\big]\ g_0(u)g_0(u_1)\cdots g_0(u_n)\ du\ du_1\cdots du_n.$$

Note that the expectation here is an expectation with respect to $u_1,\cdots,u_n$. Although $g^{(2)}_{\lambda_n}$ is not related to $u_1,\cdots,u_n$, $[ABCD]^c$ is related to $u_1,\cdots,u_n$. By $(A.9)$-$(A.11)$ and that $\int |g^{(2)}_{\lambda_n}|du \le \int |g^{(2)}_0|du$, we thus have $E(I_5) \to o_p(1)$. In the following, we will show $I_4 \to o_p(1)$ by showing that

$$\int_{[ABCD]} \left(\frac{g^{(2)}_{e\lambda_n}}{g_{e\lambda_n}} - \frac{g^{(2)}_{\lambda_n}}{g_{\lambda_n}}\right)^2 g_{\lambda_n} du \to o_p(1). \tag{$B.10$}$$

$(B.10)$ can be bounded by

$$\int_{[ABCD]} \left(\frac{g^{(2)}_{e\lambda_n}}{g_{e\lambda_n}} - \frac{g^{(2)}_{\lambda_n}}{g_{\lambda_n}}\right)^2 g_{\lambda_n} du \le$$

$$2\int_{[ABCD]} \left\{\left(\frac{g^{(2)}_{e\lambda_n} - g^{(2)}_{\lambda_n}}{g_{\lambda_n}}\right)^2 + {b_n}^2\left(\frac{g_{e\lambda_n} - g_{\lambda_n}}{g_{\lambda_n}}\right)^2\right\} g_{\lambda_n} du \tag{$B.11$}$$

The expectation of the right hand side of $(B.11)$ can be bounded by

$$2\int_{[ABCD]} g^{-1}_{\lambda_n} E\left(g^{(2)}_{e\lambda_n} - g^{(2)}_{\lambda_n}\right)^2 du + 2\int_{[ABCD]} {b_n}^2 g^{-1}_{\lambda_n} E\left(g_{e\lambda_n} - g_{\lambda_n}\right)^2 du. \tag{$B.12$}$$

Again, the expectation here is an expectation with respect to $u_1,\cdots,u_n$. By $(A.8)$, we find that $(B.12) = o(1)$. Since the right hand side of $(B.11)$ is always nonnegative, then we have that $I_4 \to o_p(1)$.

Result 3:

If $\lambda_n b_n \to 0$, then

$$\int \left|\frac{g^{(2)}_{n\lambda_n}}{g_{n\lambda_n}}\right| \left|g_{\lambda_n} - g_0\right| du \to o_p(1). \tag{$B.13$}$$

Proof: Bickel (1982), Lemma 6.3 showed that

$$\int \left(\sqrt{g_{\lambda_n}} - \sqrt{g_0}\right)^2 du \to o(\lambda_n^2).$$

We bound $(B.13)$ by

$$\int b_n \left|(\sqrt{g_{\lambda_n}} + \sqrt{g_0})(\sqrt{g_{\lambda_n}} - \sqrt{g_0})\right| du \to o_p(1).$$

Result 4:

If $\lambda_n \to 0$, then

$$\int |g^{(2)}_{\lambda_n} - g^{(2)}_0| du \to o_p(1). \tag{B.14}$$

Proof: Apply $(A.9) - (A.11)$ and that $\int |g^{(2)}_{\lambda_n}| du \leq \int |g^{(2)}_0| du$, and use the similar calculation as in $(A.15)$. This result follows by $\boldsymbol{L}_1$ continuity theorem and the dominated convergence theorem.

Given these four results, **Lemma B.1** follows when $\lambda_n \to 0$.

Lemma $B.2$:

$\sqrt{n}S^n(\boldsymbol{\beta}_0, g_{n\lambda_n}) - \sqrt{n}S^n(\boldsymbol{\beta}_0, g_0) \to o_p(1)$

Proof: We use $\boldsymbol{L}^2$ metric to prove this Lemma. Let

$$I_6 = \int \cdots \int \int \cdots \int \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right) \left(\frac{g'_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g'_0}{g_0} \right)(u_i) \right]$$

$$\left[\frac{1}{\sqrt{n}} \sum_{i=j}^{n} \left(\frac{\partial h(\boldsymbol{x}_j, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top} \right) \left(\frac{g'_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g'_0}{g_0} \right)(u_j) \right]$$

$$g_0(u_1) \cdots g_0(u_n) f(\boldsymbol{x}_1) \cdots f(\boldsymbol{x}_n) du_1 \cdots du_n d\boldsymbol{x}_1 \cdots d\boldsymbol{x}_n. \tag{B.15}$$

By the symmetry condition of $g_{n\lambda_n}$ and g_0, and that $g'_{n\lambda_n}$ and g'_0 are asymmetric with respect to their arguments, the cross products with $i \neq j$ drop out. Then we have

$$I_6 = \int \cdots \int \int \cdots \int \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top} \right) \left(\frac{g'_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g'_0}{g_0} \right)^2 (u_i) \right]$$

$$g_0(u_1) \cdots g_0(u_n) f(\boldsymbol{x}_1) \cdots f(\boldsymbol{x}_n) du_1 \cdots du_n d\boldsymbol{x}_1 \cdots d\boldsymbol{x}_n$$

$$= E\left(\frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top} \right) \int \cdots \int \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{g'_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g'_0}{g_0} \right)^2 (u_i) \right]$$

$$g_0(u_1) \cdots g_0(u_n) du_1 \cdots du_n.$$

Bickel (1982) showed that

$$\int \cdots \int \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{g'_{n\lambda_n}}{g_{n\lambda_n}} - \frac{g'_0}{g_0} \right)^2 (u_i) \right]$$

$$g_0(u_1) \cdots g_0(u_n) du_1 \cdots du_n \to o(1).$$

We then have

$$I_6 \to o(1). \tag{B.16}$$

Lemma $B.3$:

Given the regularity conditions for the asymptotic normality property of MLE, we then have $\sqrt{n}\bar{S}_n(\boldsymbol{\beta}_0, g_0) \xrightarrow{d} N(0, I(\boldsymbol{\beta}_0, g_0))$, *where* $I(\boldsymbol{\beta}_0, g_0)$ *is the information matrix of* $\boldsymbol{\beta}_0$ *under* g_0.

Proof: See Davidson and MacKinnon (1993), Chapter 8.
Given these three results, we conclude that
$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, I^{-1}(\beta_0, g_0))$.

Appendix C

Proof of Theorem 5.1:
Lemma C.1

$$-\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial h_1(x_i,\hat{\theta}_n)}{\partial\theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial\theta})\right]\frac{\tilde{g}'_{n\lambda_n}}{\tilde{g}_{n\lambda_n}}(u_i)\right\}$$

$$+\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial h_1(x_i,\theta_0)}{\partial\theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial\theta})\right]\frac{g'_0}{g_0}(u_i)\right\} \to o_p(1).$$

Proof:
Notice that, in this case, we do not assume the symmetry condition of g_0, however, we find that $\frac{\partial h_1(x_i,\theta_0)}{\partial\theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial\theta})$ has zero mean and is independent of $\left\{\frac{\tilde{g}'_{n\lambda_n}}{\tilde{g}_{n\lambda_n}} - \frac{g'_0}{g_0}\right\}(u_i)$. Let

$$I_7 = E_0\left(n^{-1/2}\sum_{i=1}^{n}\left[\frac{\partial h_1(x_i,\theta_0)}{\partial\theta} - E(\frac{\partial h_1(x,\theta_0)}{\partial\theta})\right]\left\{\frac{\tilde{g}'_{n\lambda_n}}{\tilde{g}_{n\lambda_n}} - \frac{g'_0}{g_0}\right\}(u_i)\right)^2, \qquad (C.1)$$

then the cross products drop out.

Using the same calculation in Bickel (1982), we have

$$I_7 \to o_p(1),$$

and Theorem 5.1 simply follows.

Table 1. Linear model, t3 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.0750	0.0757	0.9492-1.0469	0.9492-1.0527
$\rho = 8$	AMLE	0.0779	0.0775	0.9498-1.0488	0.9493-1.0535
$\lambda = 0.4$	SMLE	0.0749	0.0759	0.9492-1.0469	0.9492-1.0518
$\rho = 12$	AMLE	0.0779	0.0775	0.9498-1.0488	0.9493-1.0535
$\lambda = 0.5$	SMLE	0.0750	0.0758	0.9502-1.0469	0.9502-1.0508
$\rho = 8$	AMLE	0.0793	0.0791	0.9483-1.0488	0.9466-1.0534
$\lambda = 0.5$	SMLE	0.0747	0.0759	0.9492-1.0469	0.9512-1.0508
$\rho = 12$	AMLE	0.0793	0.0791	0.9483-1.0488	0.9466-1.0534
	OLSE	0.1024	0.1040	0.9295-1.0640	0.9387-1.0698

Table 2. Linear model, t5 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.0919	0.0920	0.9356-1.0625	0.9336-1.0606
$\rho = 8$	AMLE	0.0939	0.0949	0.9326-1.0638	0.9302-1.0618
$\lambda = 0.4$	SMLE	0.0918	0.0918	0.9356-1.0625	0.9336-1.0606
$\rho = 12$	AMLE	0.0939	0.0949	0.9326-1.0638	0.9302-1.0618
$\lambda = 0.5$	SMLE	0.0913	0.0916	0.9375-1.0615	0.9336-1.0625
$\rho = 8$	AMLE	0.0928	0.0942	0.9320-1.0623	0.9322-1.0627
$\lambda = 0.5$	SMLE	0.0912	0.0916	0.9375-1.0615	0.9336-1.0625
$\rho = 12$	AMLE	0.0928	0.0942	0.9320-1.0623	0.9322-1.0627
	OLSE	0.1003	0.0997	0.9296-1.0667	0.9281-1.0660

Table 3. Linear model, t8 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	interquartile range β_0	interquartile range β_1
$\lambda = 0.4$	SMLE	0.0975	0.0972	0.9356-1.0664	0.9336-1.0625
$\rho = 8$	AMLE	0.1008	0.1004	0.9310-1.0706	0.9328-1.0661
$\lambda = 0.4$	SMLE	0.0975	0.0972	0.9356-1.0664	0.9336-1.0625
$\rho = 12$	AMLE	0.1008	0.1004	0.9310-1.0706	0.9328-1.0661
$\lambda = 0.5$	SMLE	0.0968	0.0965	0.9375-1.0664	0.9336-1.0625
$\rho = 8$	AMLE	0.0994	0.0990	0.9343-1.0696	0.9307-1.0645
$\lambda = 0.5$	SMLE	0.0968	0.0965	0.9375-1.0664	0.9336-1.0625
$\rho = 12$	AMLE	0.0994	0.0990	0.9343-1.0696	0.9307-1.0645
	OLSE	0.0997	0.1009	0.9362-1.0666	0.9320-1.0640

Table 4. Linear model, Normal errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.1046	0.1017	0.9297-1.0762	0.9356-1.0625
$\rho = 8$	AMLE	0.1081	0.1028	0.9278-1.0749	0.9306-1.0640
$\lambda = 0.4$	SMLE	0.1046	0.0995	0.9297-1.0772	0.9356-1.0625
$\rho = 12$	AMLE	0.1081	0.1028	0.9278-1.0749	0.9306-1.0640
$\lambda = 0.5$	SMLE	0.1039	0.0986	0.9297-1.0742	0.9375-1.0625
$\rho = 8$	AMLE	0.1060	0.1005	0.9298-1.0769	0.9335-1.0646
$\lambda = 0.5$	SMLE	0.1038	0.0986	0.9297-1.0742	0.9375-1.0625
$\rho = 12$	AMLE	0.1060	0.1005	0.9298-1.0769	0.9335-1.0646
	OLSE	0.1025	0.0977	0.9315-1.0725	0.9363-1.0632

Table 5. Linear model, BN1 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.1275	0.1255	0.9219-1.0938	0.9102-1.0781
$\rho = 8$	AMLE	0.1291	0.1485	0.9263-1.0946	0.9102-1.0918
$\lambda = 0.4$	SMLE	0.1277	0.1253	0.9219-1.0938	0.9102-1.0781
$\rho = 12$	AMLE	0.1291	0.1485	0.9263-1.0946	0.9102-1.0918
$\lambda = 0.5$	SMLE	0.1175	0.1177	0.9258-1.0898	0.9180-1.0742
$\rho = 8$	AMLE	0.1328	0.1420	0.9193-1.0970	0.9053-1.0902
$\lambda = 0.5$	SMLE	0.1175	0.1177	0.9258-1.0898	0.9180-1.0742
$\rho = 12$	AMLE	0.1328	0.1420	0.9193-1.0970	0.9053-1.0902
	OLSE	0.3162	0.3208	0.7779-1.2031	0.7838-1.2014

Table 6. Linear model, BN2 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.1457	0.1455	0.9063-1.1016	0.9063-1.0898
$\rho = 8$	AMLE	0.1479	0.1549	0.9034-1.0977	0.8921-1.0908
$\lambda = 0.4$	SMLE	0.1458	0.1455	0.9063-1.0996	0.9063-1.0889
$\rho = 12$	AMLE	0.1479	0.1549	0.9034-1.0977	0.8921-1.0908
$\lambda = 0.5$	SMLE	0.1376	0.1388	0.9063-1.0938	0.9082-1.0859
$\rho = 8$	AMLE	0.1474	0.1578	0.9023-1.0981	0.8895-1.0940
$\lambda = 0.5$	SMLE	0.1376	0.1388	0.9063-1.0938	0.9082-1.0859
$\rho = 12$	AMLE	0.1474	0.1578	0.9023-1.0981	0.88951-1.0940
	OLSE	0.2282	0.2286	0.8510-1.1621	0.8451-1.1639

Table 7. Linear model, BN3 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.1469	0.1455	0.9033-1.1016	0.9043-1.1094
$\rho = 8$	AMLE	0.1530	0.1534	0.8983-1.1072	0.8974-1.1117
$\lambda = 0.4$	SMLE	0.1453	0.1451	0.9033-1.1016	0.9043-1.1094
$\rho = 12$	AMLE	0.1530	0.1534	0.8983-1.1072	0.8974-1.1117
$\lambda = 0.5$	SMLE	0.1434	0.1438	0.9063-1.1016	0.9063-1.1094
$\rho = 8$	AMLE	0.1480	0.1492	0.9025-1.1061	0.9016-1.1117
$\lambda = 0.5$	SMLE	0.1434	0.1438	0.9063-1.1016	0.9063-1.1094
$\rho = 12$	AMLE	0.1480	0.1492	0.9025-1.1061	0.9016-1.1117
	OLSE	0.1457	0.1463	0.9041-1.1007	0.9072-1.1113

Table 8. Linear model, LN1 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.0998	0.0340	0.9990-1.1100	0.9766-1.0205
$\rho = 8$	AMLE	0.2146	0.1027	1.1287-1.2529	0.9328-1.0763
$\lambda = 0.4$	SMLE	0.1006	0.0341	0.9990-1.1100	0.9766-1.0205
$\rho = 12$	AMLE	0.2146	0.1017	1.1287-1.2529	0.9328-1.0763
$\lambda = 0.5$	SMLE	0.1047	0.0370	1.0046-1.1169	0.9756-1.0227
$\rho = 8$	AMLE	0.2501	0.1198	1.1653-1.2901	0.9246-1.0829
$\lambda = 0.5$	SMLE	0.1055	0.0371	1.0046-1.1169	0.9756-1.0227
$\rho = 12$	AMLE	0.2501	0.1198	1.1653-1.2901	0.9246-1.0829
	OLSE	0.0798	0.0825	0.9441-1.0470	0.9490-1.0494

Table 9. Linear model, LN2 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.0724	0.0377	1.0154-1.0850	0.9761-1.0256
$\rho = 8$	AMLE	0.1417	0.0460	1.0859-1.1681	0.9743-1.0354
$\lambda = 0.4$	SMLE	0.0724	0.0377	1.0154-1.0850	0.9761-1.0256
$\rho = 12$	AMLE	0.1417	0.0460	1.0859-1.1681	0.9743-1.0354
$\lambda = 0.5$	SMLE	0.0752	0.0396	1.0173-1.0879	0.9746-1.0281
$\rho = 8$	AMLE	0.1661	0.0483	1.1079-1.1952	0.9732-1.0361
$\lambda = 0.5$	SMLE	0.0752	0.0396	1.0173-1.0879	0.9746-1.0281
$\rho = 12$	AMLE	0.1661	0.0483	1.1079-1.1952	0.9732-1.0361
	OLSE	0.0475	0.0470	0.9700-1.0325	0.9694-1.0327

Table 10. Linear model, LN3 errors (n=100)

parameter		Root MSE β_0	Root MSE β_1	Interquartile range β_0	Interquartile range β_1
$\lambda = 0.4$	SMLE	0.0084	0.0086	0.9966-1.0057	0.9944-1.0039
$\rho = 8$	AMLE	0.0207	0.0101	1.0012-1.0237	0.9932-1.0064
$\lambda = 0.4$	SMLE	0.0096	0.0097	0.9951-1.0068	0.9941-1.0059
$\rho = 12$	AMLE	0.0207	0.0101	1.0012-1.0237	0.9932-1.0064
$\lambda = 0.5$	SMLE	0.0096	0.0097	0.9941-1.0068	0.9934-1.0059
$\rho = 8$	AMLE	0.0211	0.0101	1.0014-1.0241	0.9932-1.0064
$\lambda = 0.5$	SMLE	0.0097	0.0097	0.9941-1.0068	0.9932-1.0059
$\rho = 12$	AMLE	0.0211	0.0101	1.0014-1.0241	0.9932-1.0064
	OLSE	0.0098	0.0100	0.9939-1.0067	0.9933-1.0064

Table 11. Ratio of root MSE of SMLE to root MSE of OLSE

	t3	t5	t8	N	BN1	BN2	BN3	LN1	LN2	LN3
β_0	0.7295	0.9089	0.9715	1.0136	0.3716	0.6031	0.7784	1.3227	1.5808	0.9847
β_1	0.7296	0.9183	0.9564	1.0084	0.3670	0.6071	0.9828	0.4500	0.8413	0.9759

Table 12. Ratio of root MSE of SMLE to root MSE of AMLE

	t3	t5	t8	N	BN1	BN2	BN3	LN1	LN2	LN3
β_0	0.9425	0.9819	0.9743	0.9794	0.8849	0.9338	0.7664	0.4218	0.4525	0.4580
β_1	0.9592	0.9727	0.9741	0.9806	0.8290	0.8796	0.9637	0.3098	0.8188	0.9633

Table 13. Ratio of root MSE of AMLE to root MSE of OLSE

	t3	t5	t8	N	BN1	BN2	BN3	LN1	LN2	LN3
β_0	0.7740	0.9256	0.9971	1.0349	0.4200	0.6459	1.0156	3.1359	3.4933	2.1150
β_1	0.7607	0.9441	0.9819	1.0283	0.4427	0.6902	1.0199	1.4524	1.0274	1.0131

Table 14. Mean of estimates of β_0, log-normal errors

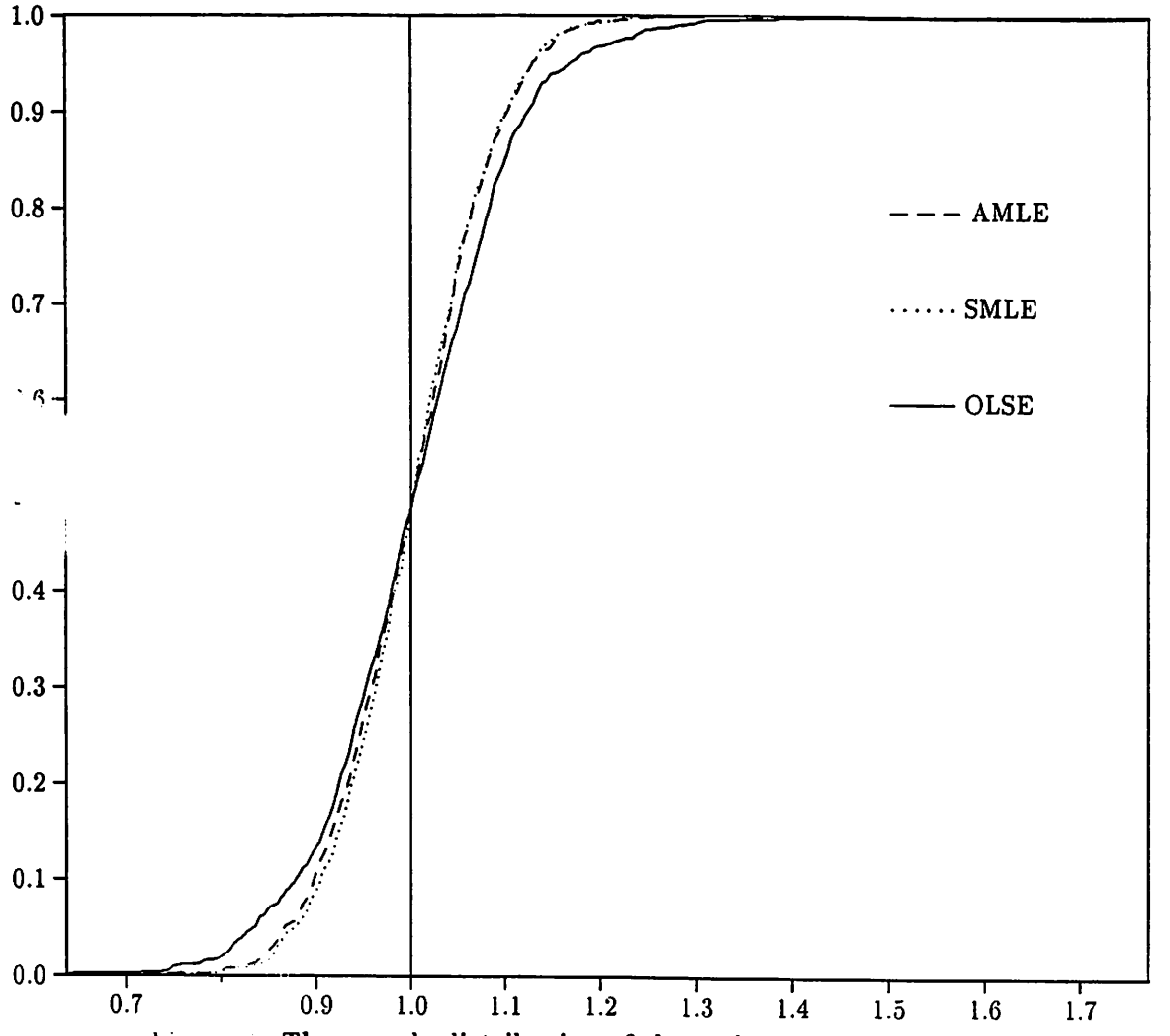
	LN1	LN2	LN3
SMLE	1.0638	1.0533	1.0007
OLSE	0.9983	1.0021	1.0002
AMLE	1.2304	1.1537	1.0131

Table 15. Linear model, Different sample size

		$n = 50$		$n = 200$	
	parameter	β_0	β_1	β_0	β_1
t3	SMLE	0.1112	0.1109	0.0517	0.0532
	AMLE	0.1180	0.1210	0.0545	0.0572
	OLSE	0.1455	0.1417	0.0707	0.0688
BN1	SMLE	0.2105	0.1984	0.0757	0.0720
	AMLE	0.2157	0.2455	0.0869	0.0864
	OLSE	0.4347	0.4638	0.2145	0.2270
LN1	SMLE	0.1341	0.0587	0.0898	0.0260
	AMLE	0.2672	0.1571	0.2453	0.0934
	OLSE	0.1154	0.1186	0.0551	0.0572

Table 16. Ratio of root MSE, Different sample size

	parameter	SMLE/OLSE β_0	SMLE/OLSE β_1	SMLE/AMLE β_0	SMLE/AMLE β_1	AMLE/OLSE β_0	AMLE/OLSE β_1
t3	n=50	0.7639	0.7826	0.9418	0.9162	0.8111	0.8541
	n=100	0.7295	0.7296	0.9425	0.9592	0.7740	0.7607
	n=200	0.7312	0.7727	0.9483	0.9298	0.7711	0.8310
BN1	n=50	0.4842	0.4278	0.9758	0.8085	0.4962	0.5292
	n=100	0.3716	0.3670	0.8849	0.8290	0.4200	0.4427
	n=200	0.3528	0.3174	0.8707	0.8339	0.4052	0.3806
LN1	n=50	1.1624	0.4950	0.5019	0.3736	2.3158	1.3248
	n=100	1.3227	0.4500	0.4218	0.3098	3.1359	1.4524
	n=200	1.6308	0.4542	0.3662	0.2782	4.4538	1.6327



. The sample distribution of the estimates of β_1. t3 case.

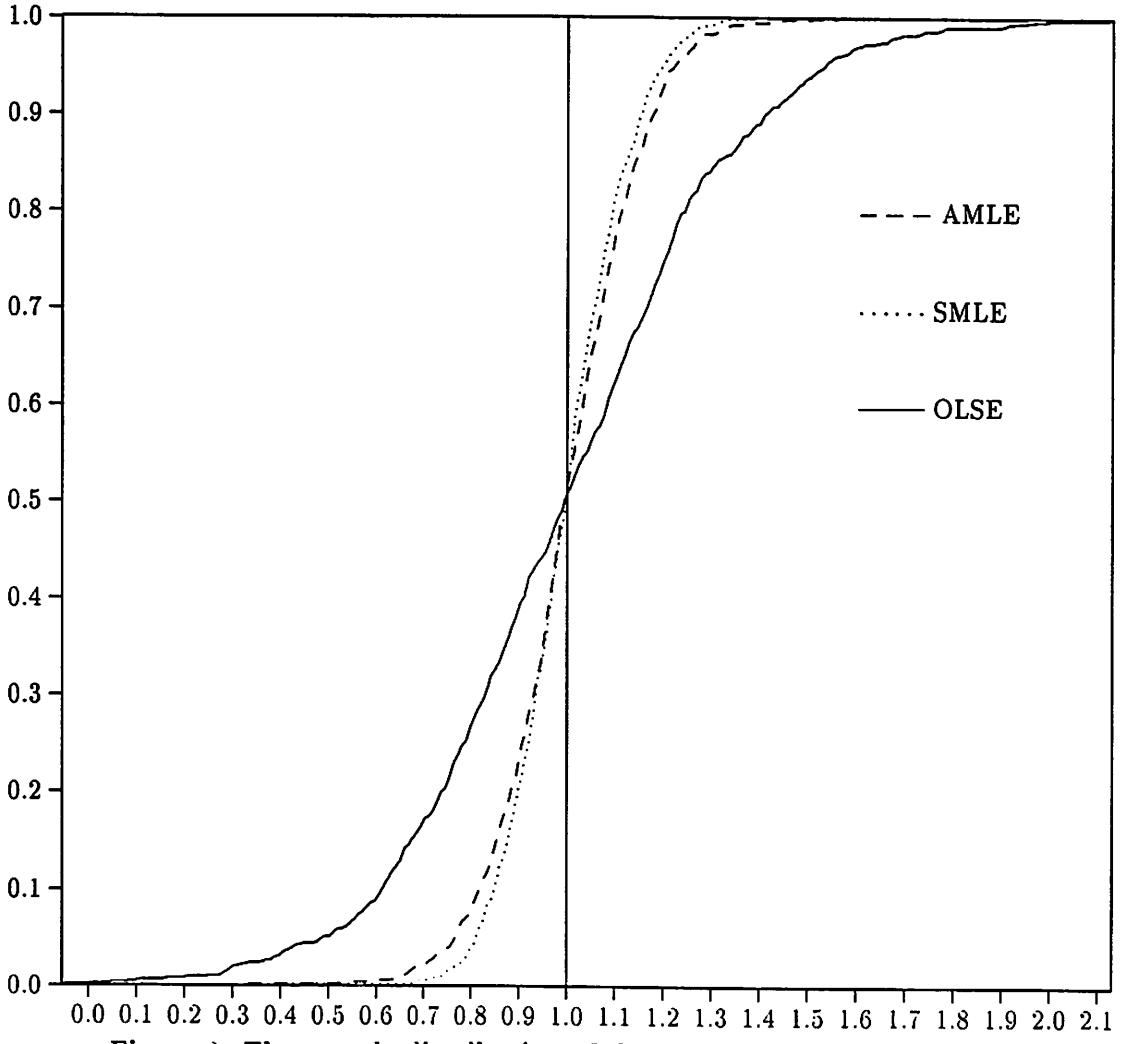


Figure 2. The sample distribution of the estimates of β_1, BN1 case.

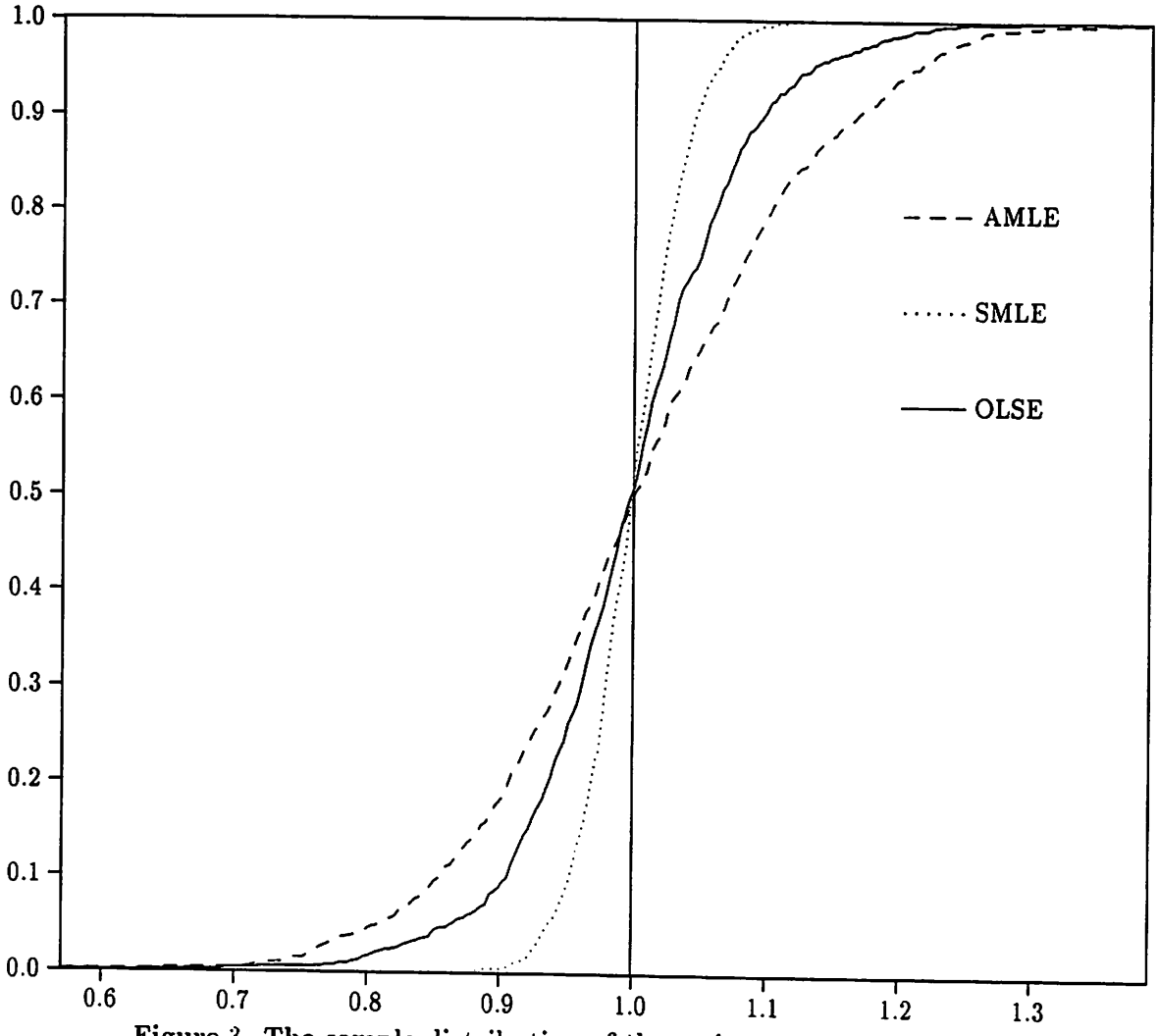


Figure 3. The sample distribution of the estimates of β_1, LN1 case.

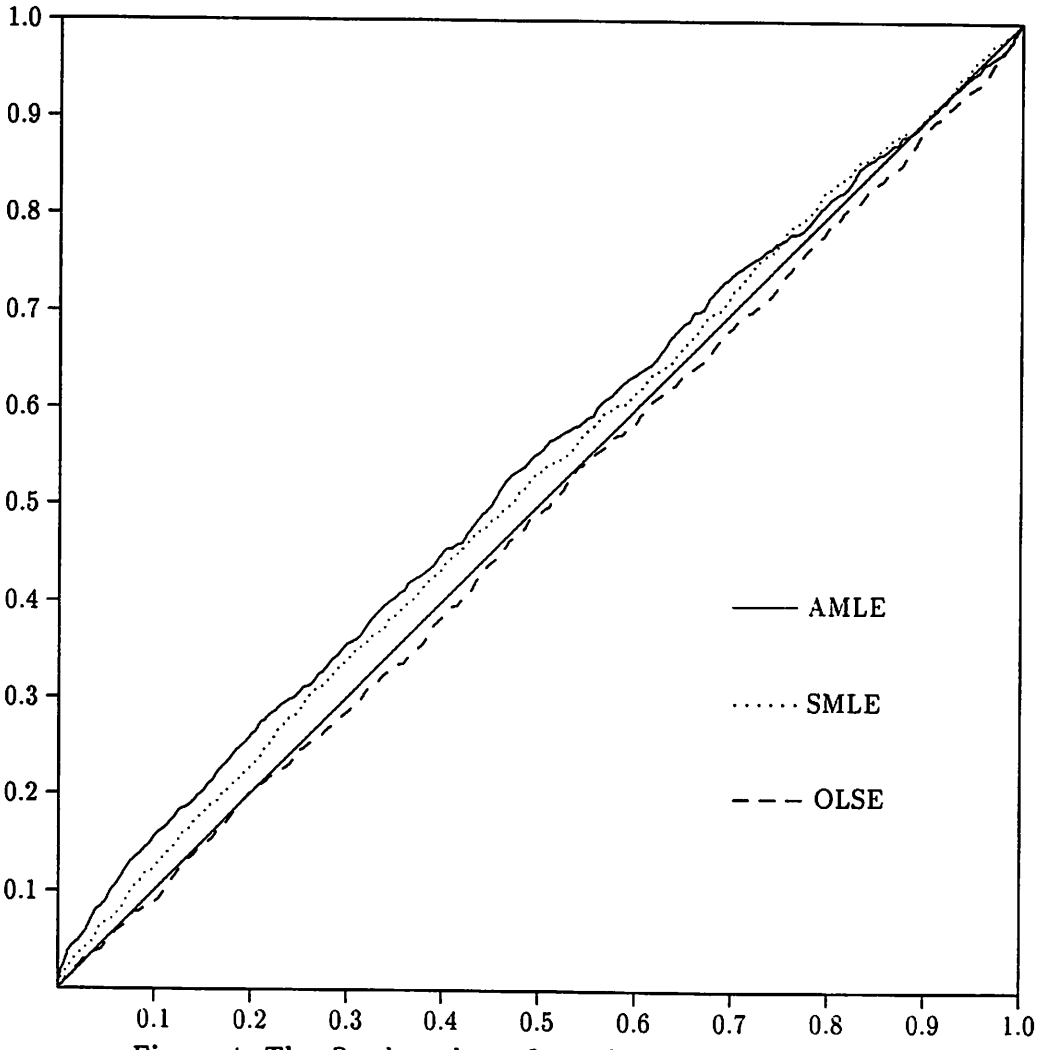


Figure 4. The P value plots of t-ratio test, t3 case.n=100.

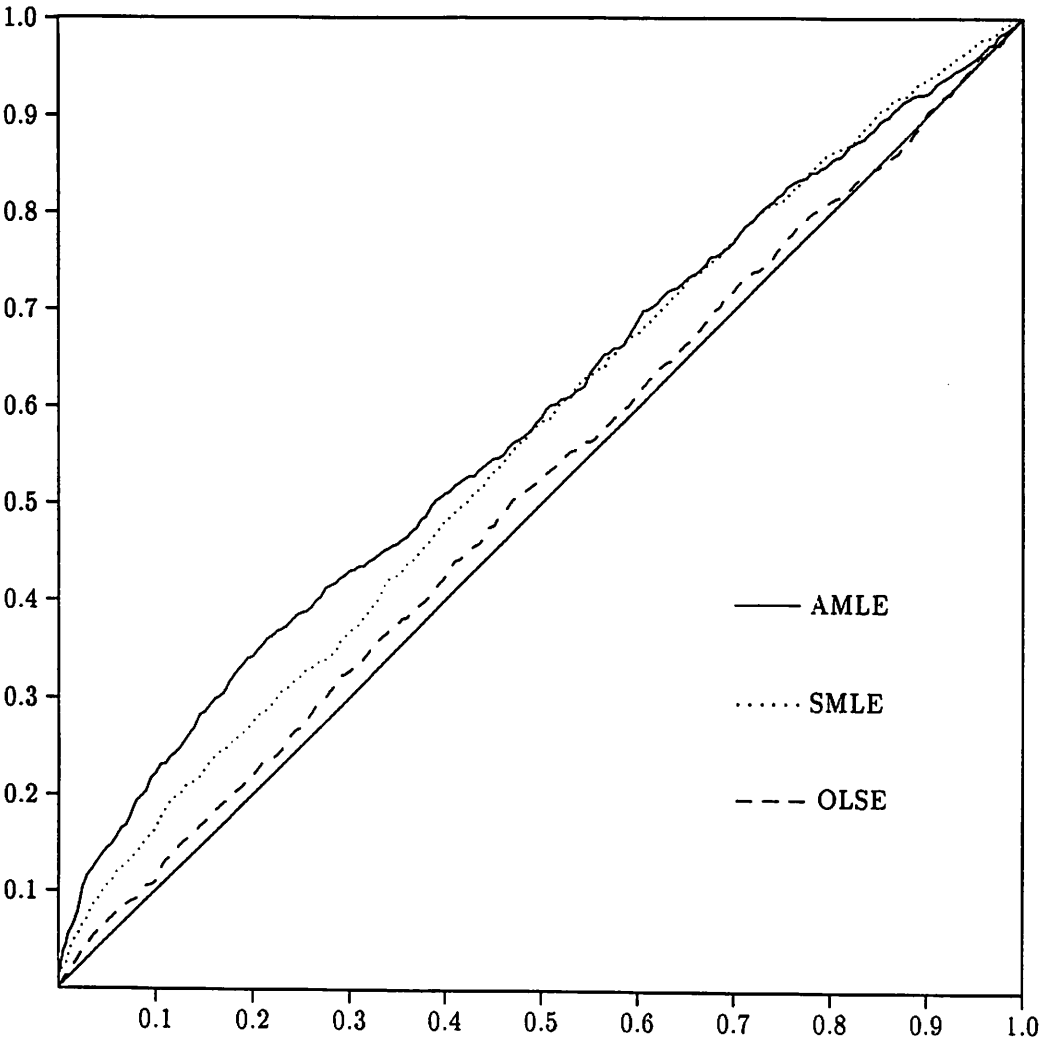


Figure 5. The P value plots of t-ratio test, BN1 case, n=100.

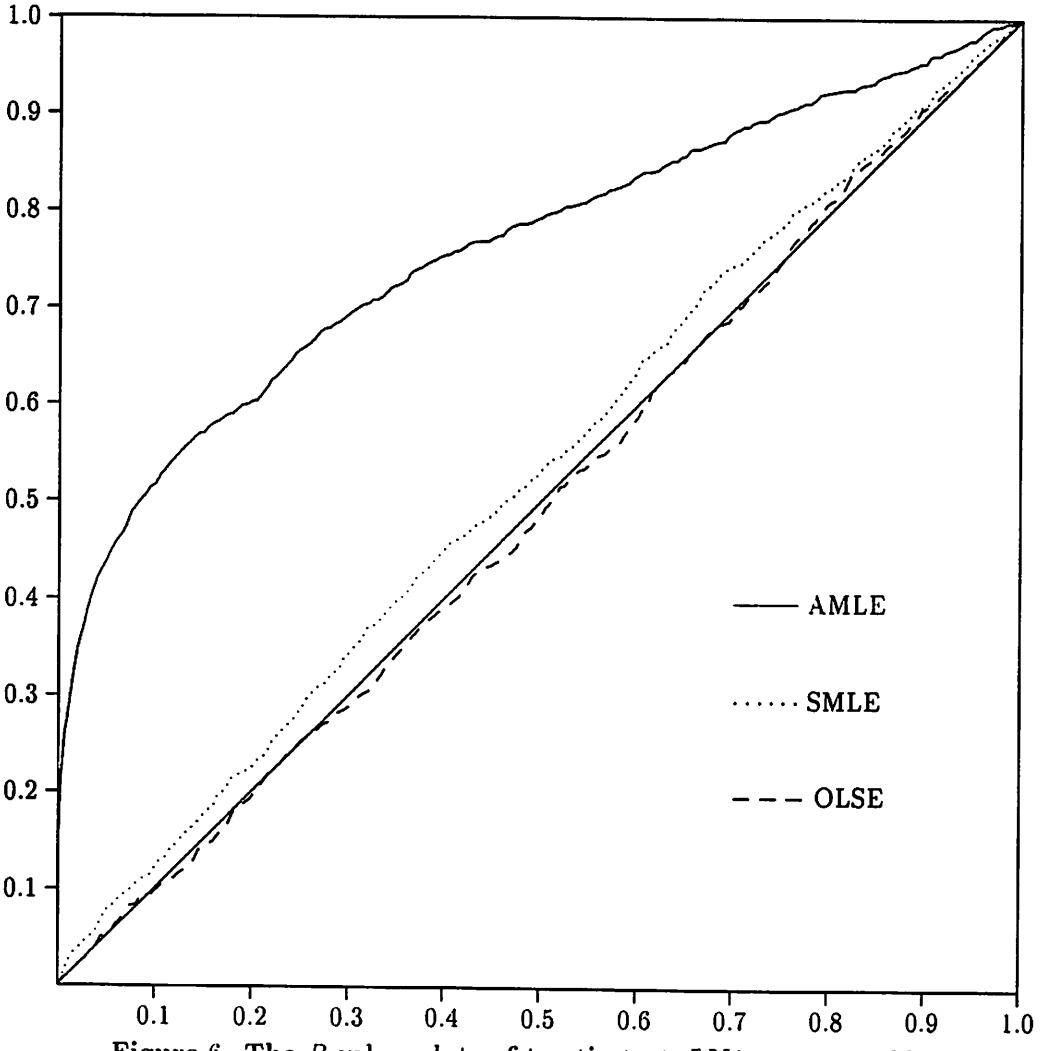


Figure 6. The P value plots of t-ratio test. LN1 case, n=100.

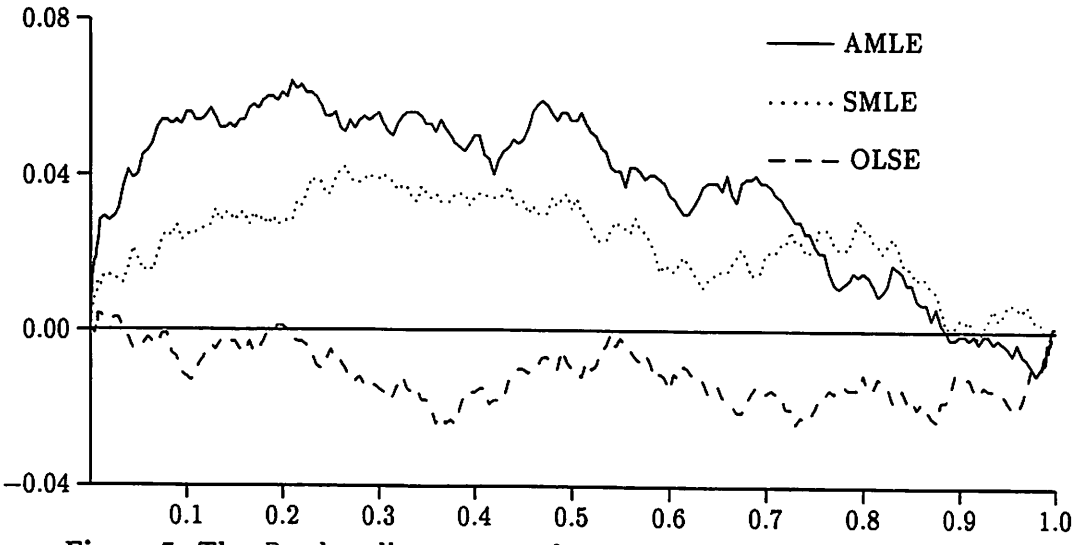


Figure 7. The P value discrepancy plots of t-ratio test, t3 case,n=100.

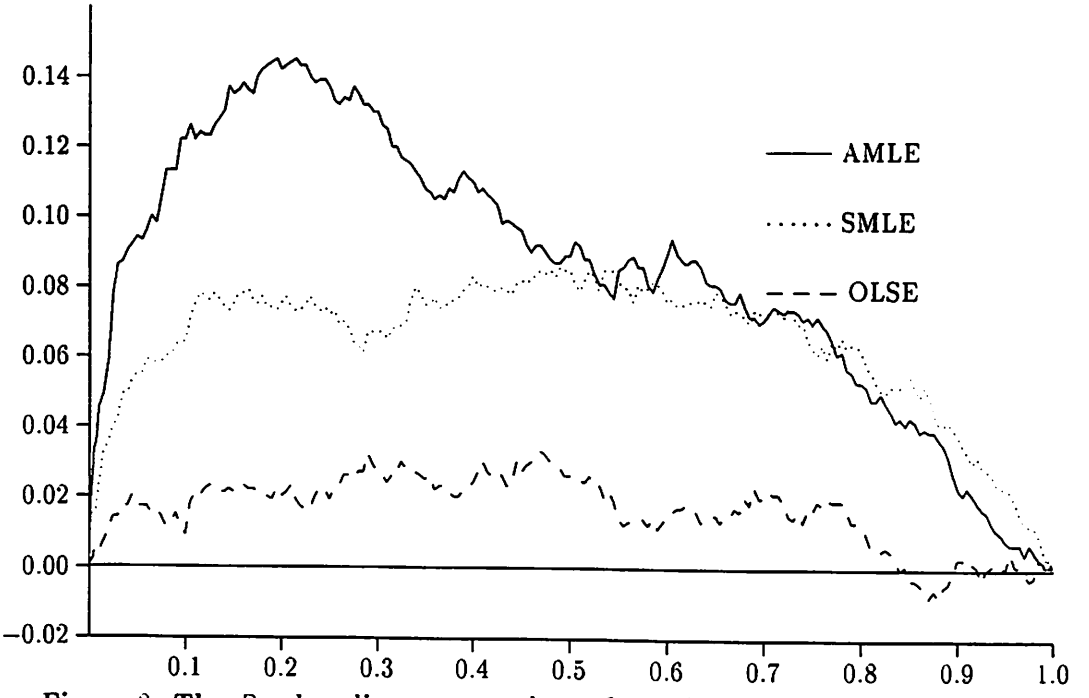


Figure 8. The P value discrepancy plots of t-ratio test, BN1 case, n=100.

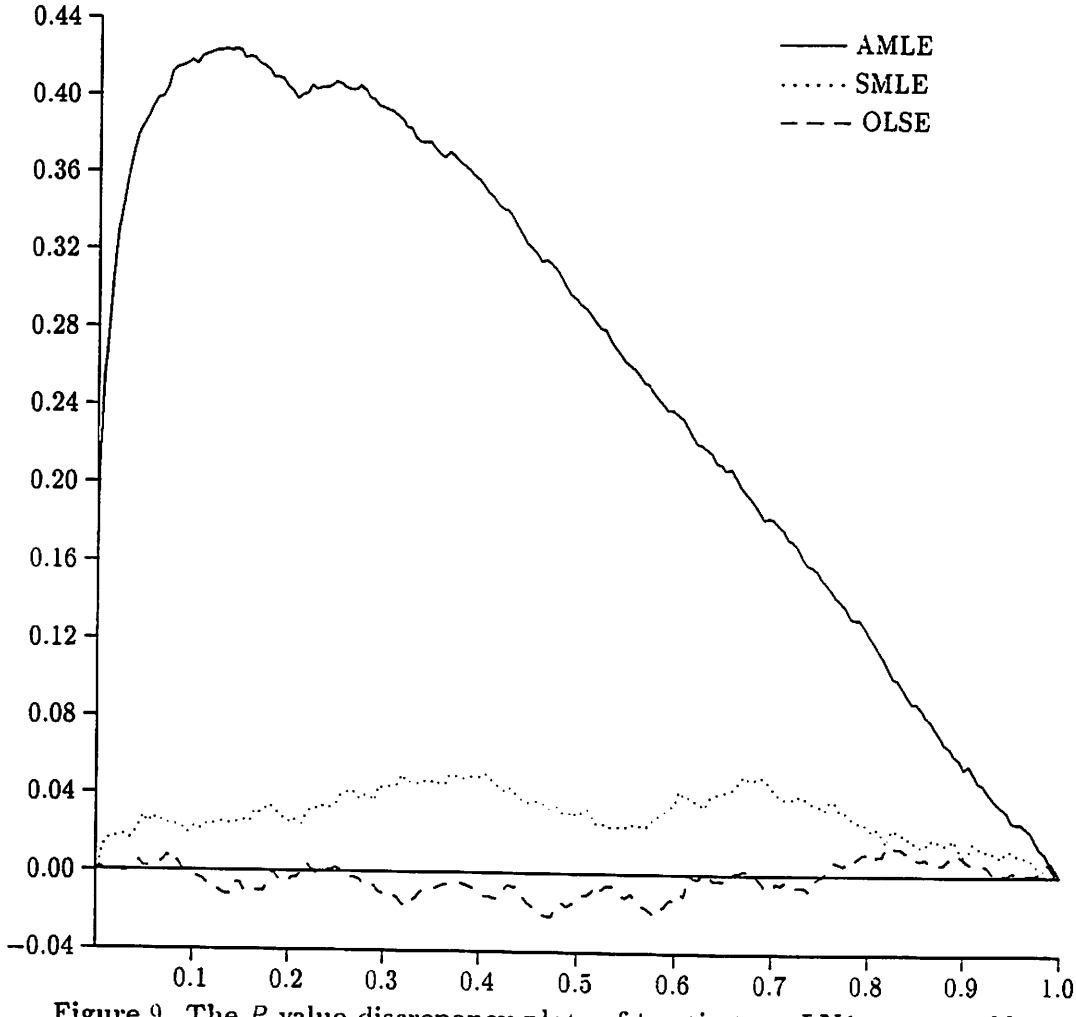


Figure 9. The P value discrepancy plots of t-ratio test. LN1 case.n=100.

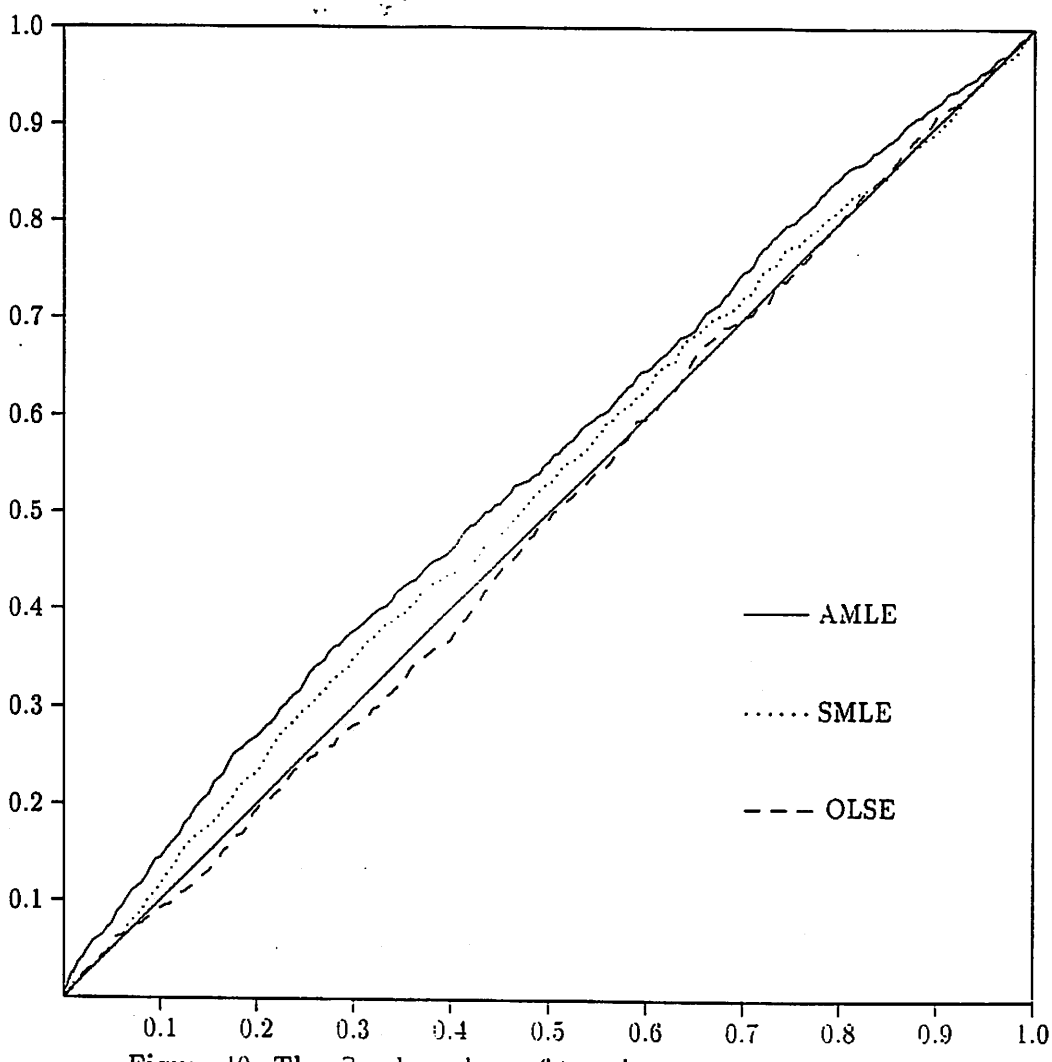


Figure 10. The P value plots of t-ratio test. t3 case. n=200.

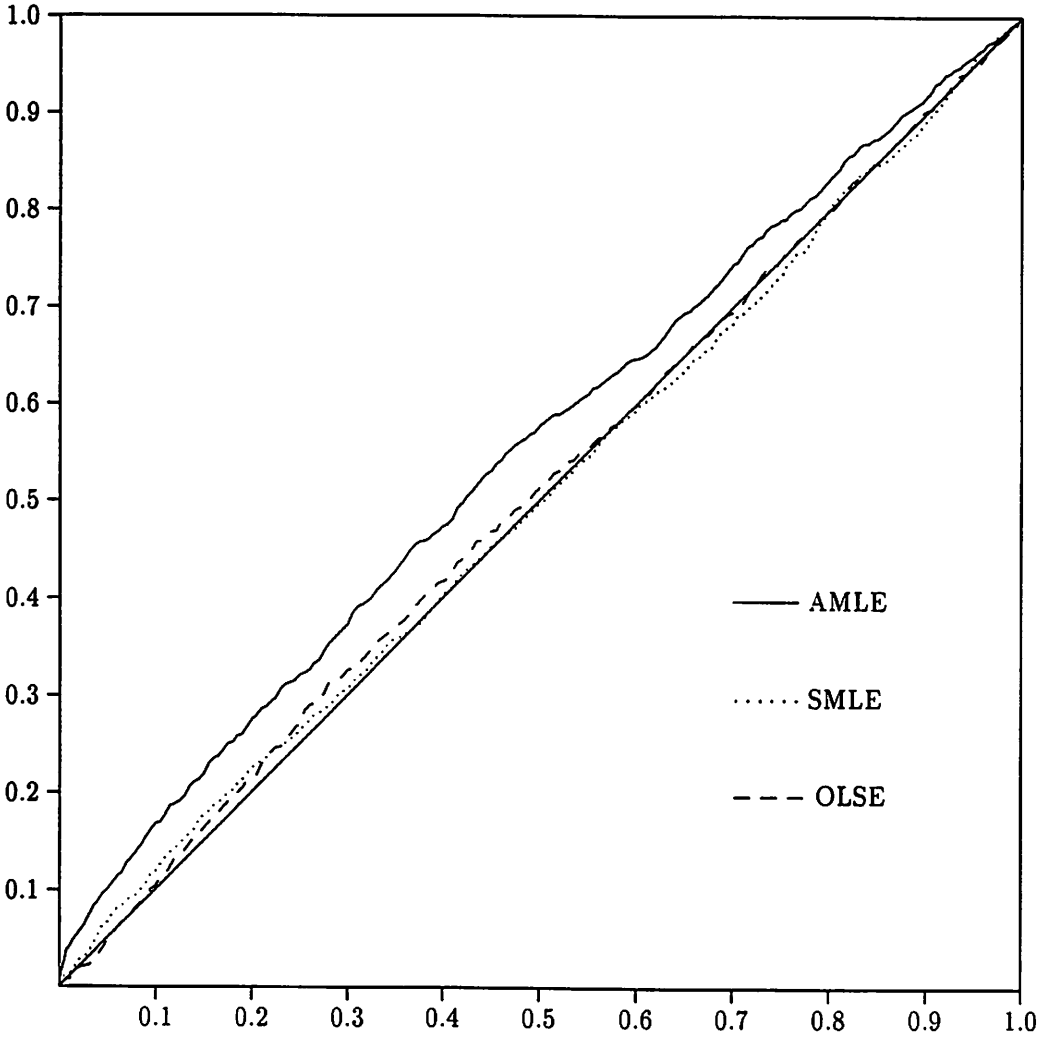


Figure 11. The P value plots of t-ratio test, BN1 case, n=200.

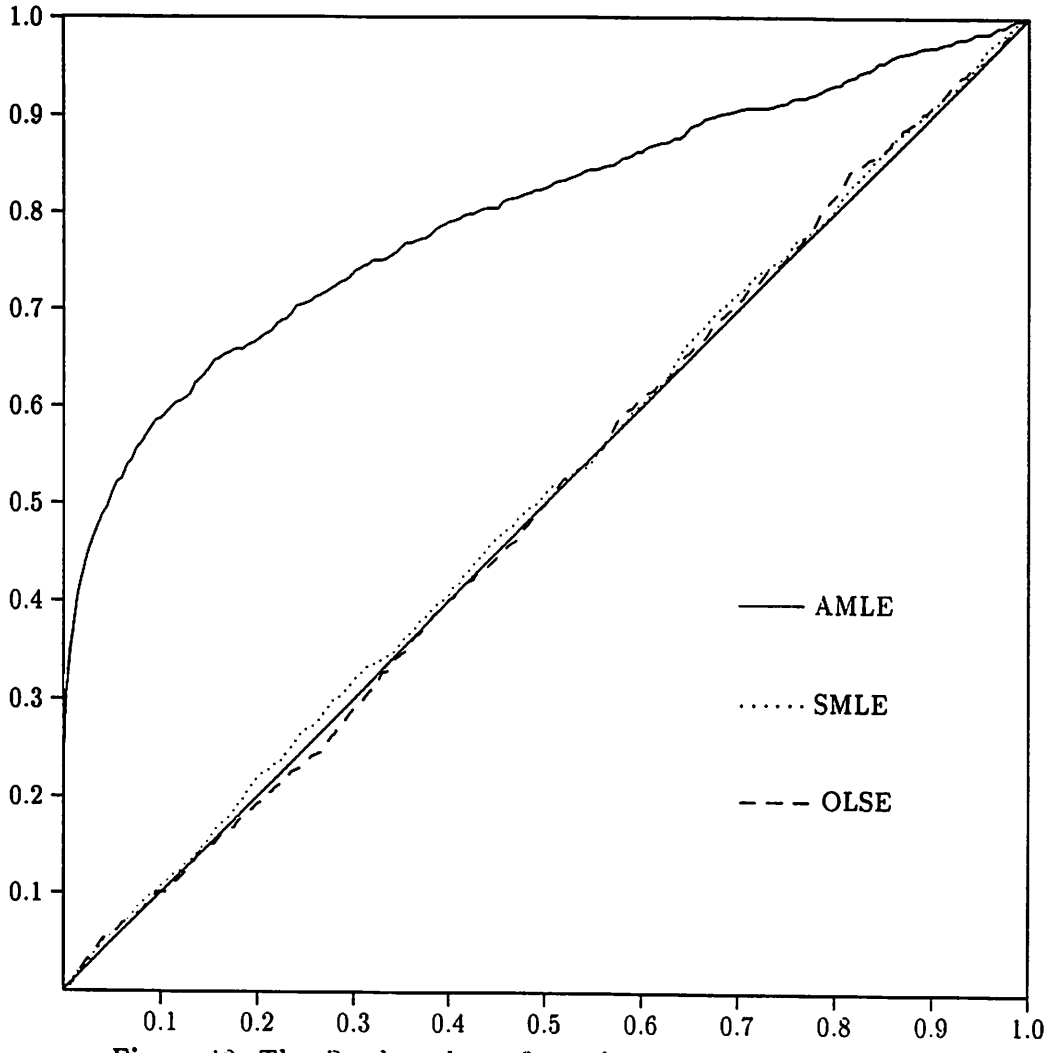


Figure 12. The P value plots of t-ratio test. LN1 case. n=200.

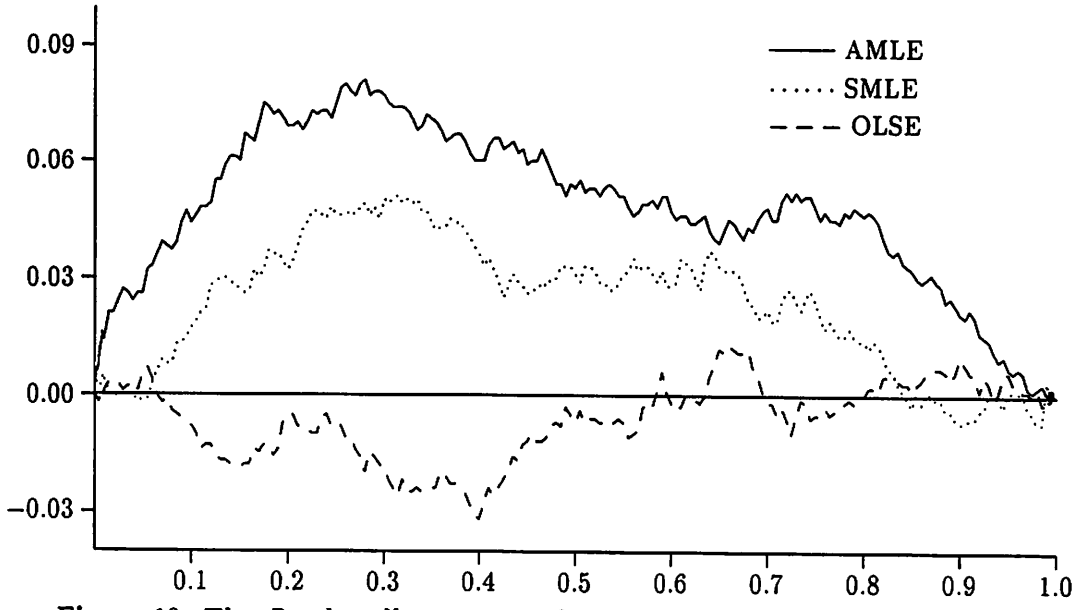


Figure 13. The P value discrepancy plots of t-ratio test, t3 case, n=200.

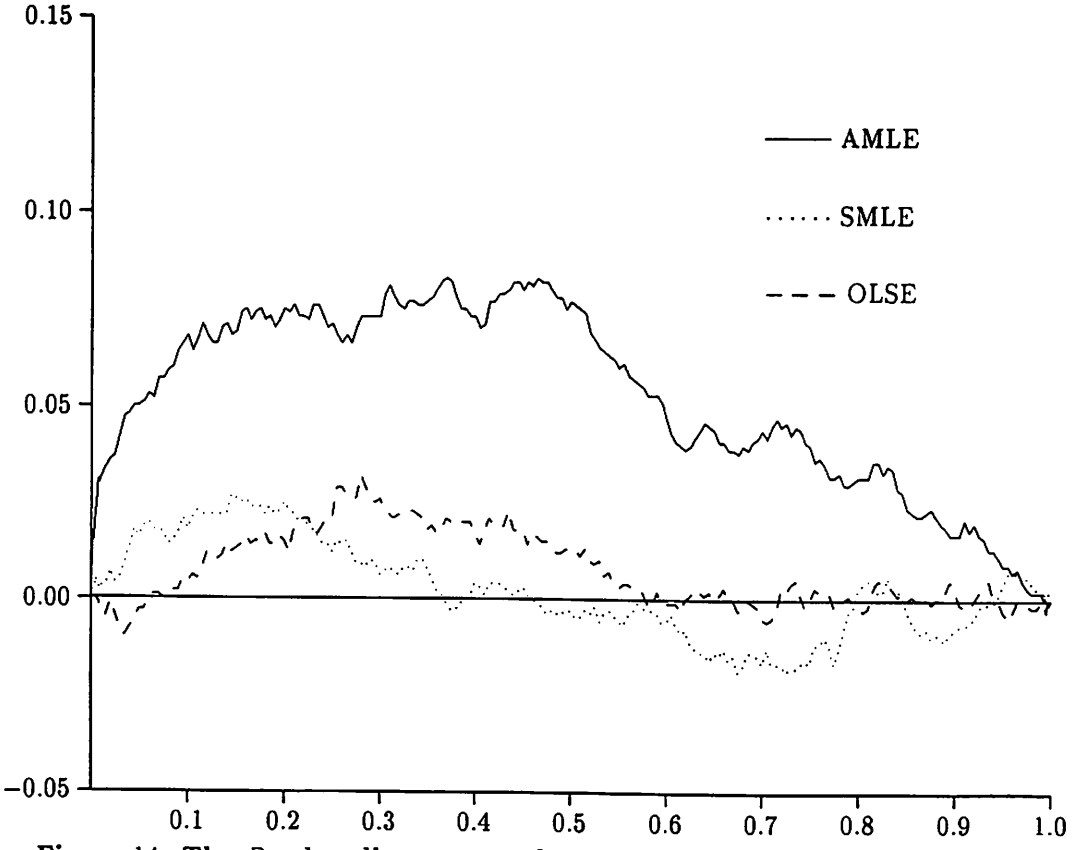


Figure 14. The P value discrepancy plots of t-ratio test, BN1 case, n=200.

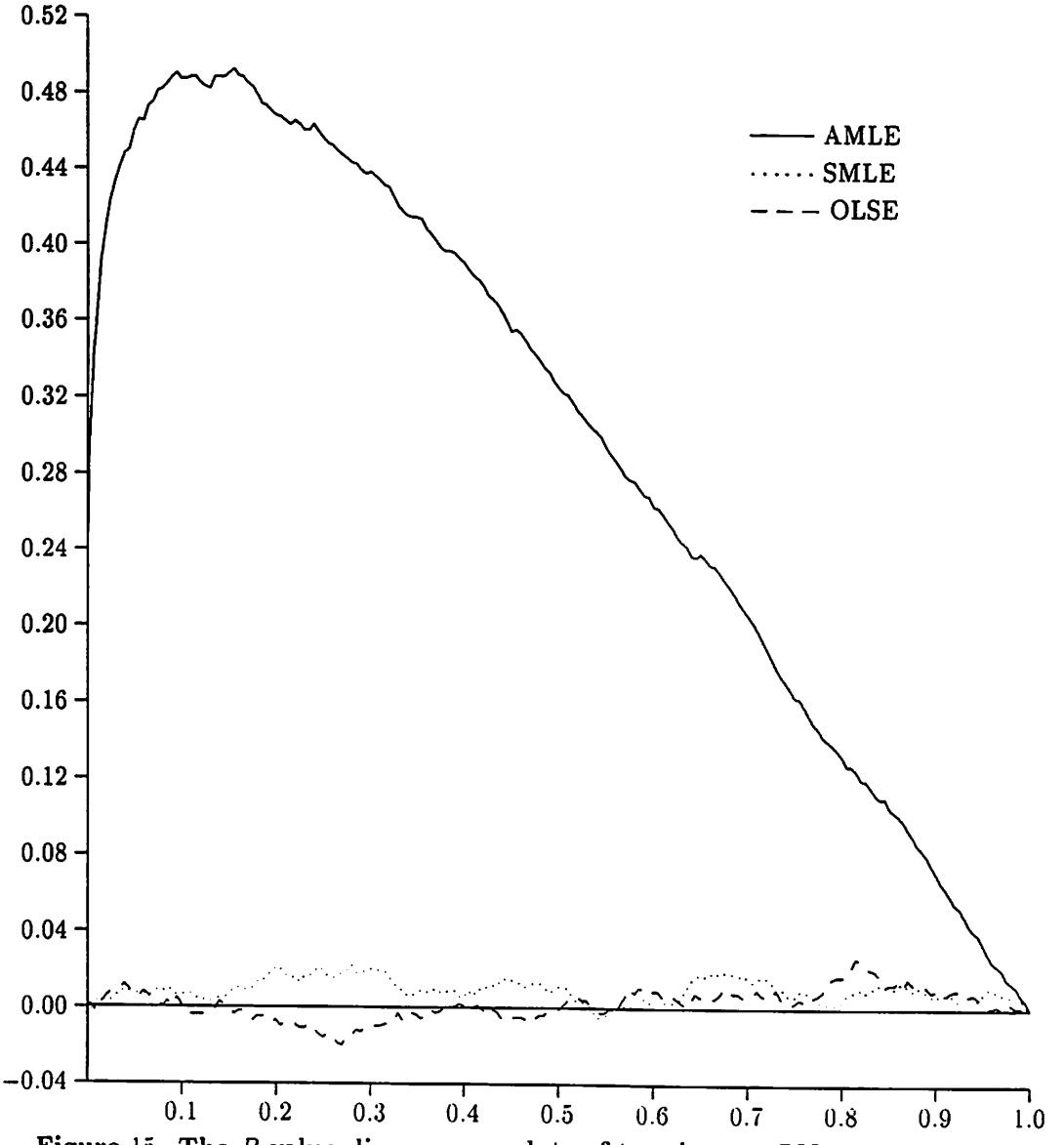


Figure 15. The P value discrepancy plots of t-ratio test, LN1 case, n=200.

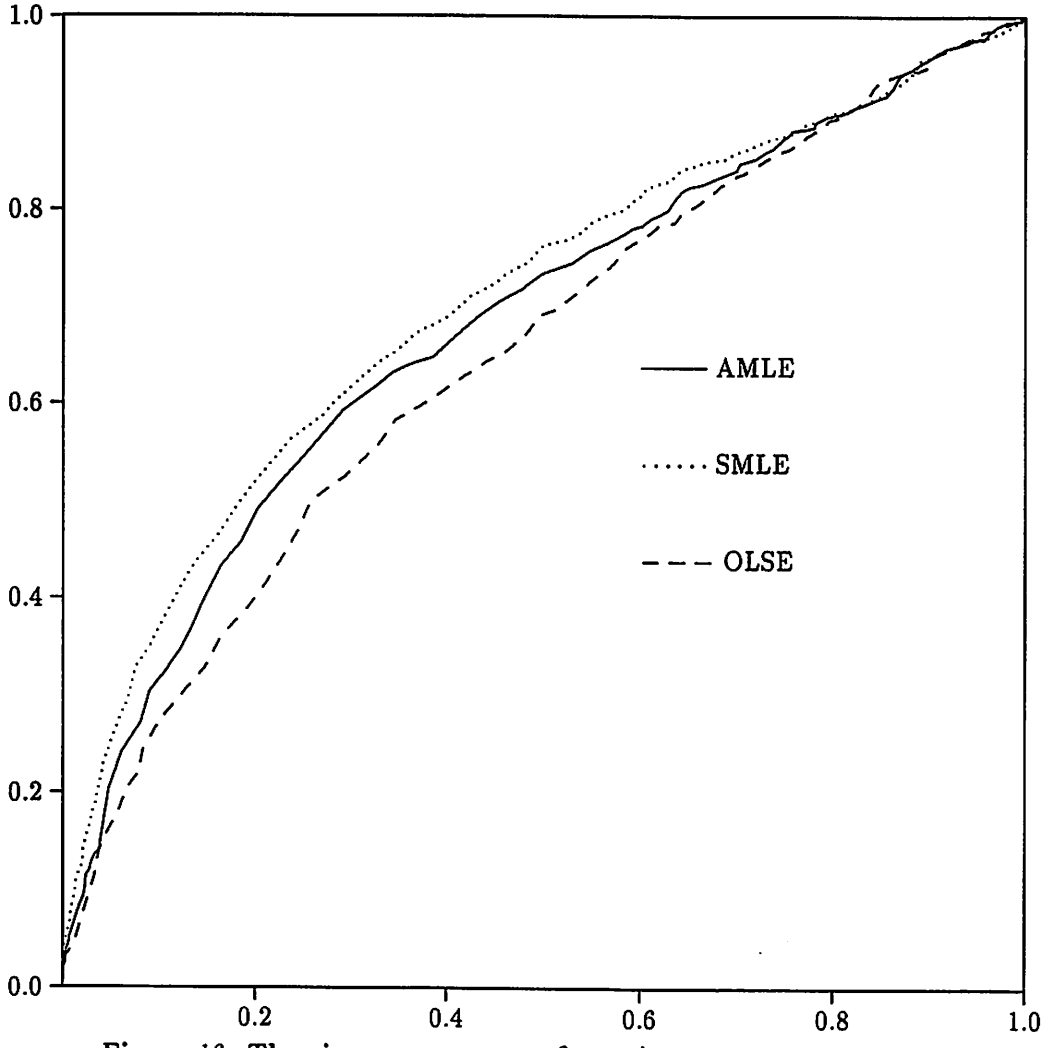


Figure 16. The size-power curve of t-ratio test, t3 case, n=100.

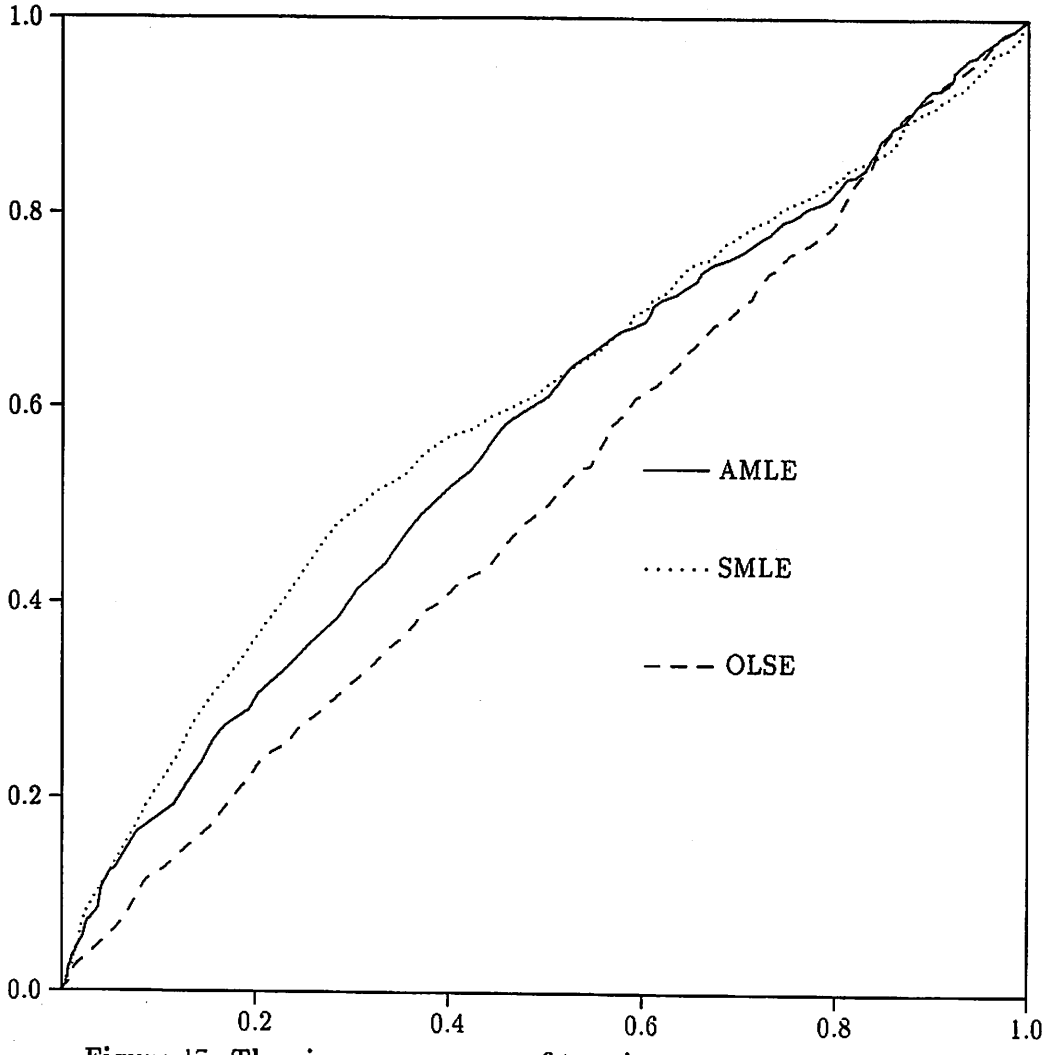


Figure 17. The size-power curve of t-ratio test, BN1 case. n=100.

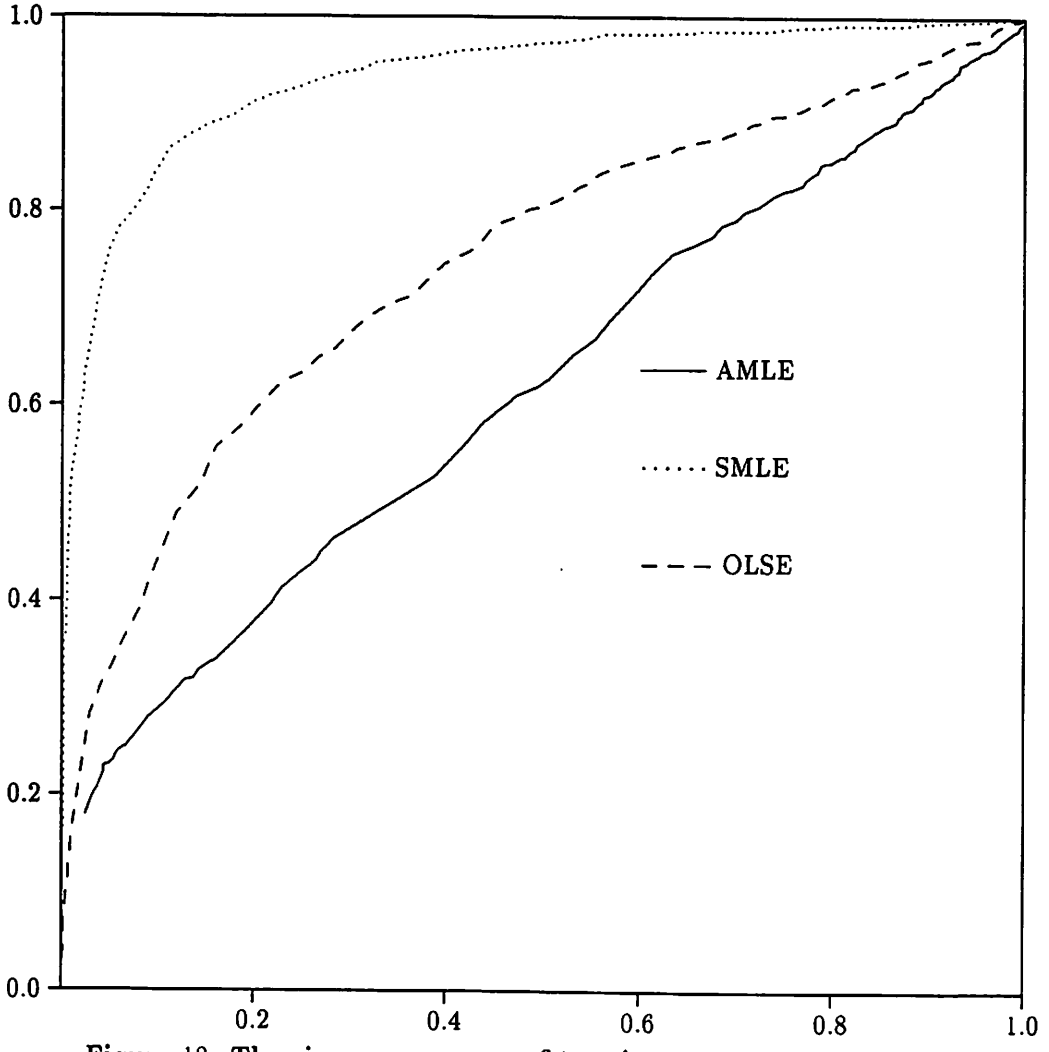


Figure 18. The size-power curve of t-ratio test, LN1 case, n=100.